The Dissertation Committee for Michael David Cave Certifies that this is the approved version of the following dissertation:

Scalable Voltage Reference for Ultra Deep Submicron Technologies

Committee:

John Davis, Supervisor

Francis Bostick

Jack Lee

John Pearce

Randall Geiger

Scalable Voltage Reference for Ultra Deep Submicron Technologies

by

Michael David Cave, B.S.E.E.; M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2005

Dedication

I dedicate this work to my family.

Acknowledgements

I would like to thank Dr. Davis and Dr. Bostick for their enthusiasm for my research. They generously supported me in pursuing my academic goals on a part-time basis for so many years.

I would like to thank Dr. Geiger from Iowa State University, who is a member of my dissertation committee and who was a significant contributor to my research. I appreciate all of the time and effort that he provided in guiding me through this experience.

Peter Linden and Dong Wong from Silvaco donated their IC development tools and resources to complete this dissertation. Their response and dedication in supporting me through this project is greatly appreciated. I would like to give them special thanks.

I would also like to thank the other members of my committee, Dr. Lee and Dr. Pearce. Their support and feedback is greatly appreciated.

Scalable Voltage Reference for Ultra Deep Submicron Technologies

Publication No._____

Michael David Cave, Ph.D.

The University of Texas at Austin, 2005

Supervisor: John H. Davis

A CMOS voltage reference architecture is shown that operates at a power supply that is equal to two threshold voltages. The circuit designed includes no more than two devices stacked between V_{dd} and V_{ss}, thus allowing for the circuit to scale with device sizing (technology). The circuit is demonstrated in a 0.18μ and 0.09μ process that operates at a power supply voltage of 0.9V and 0.4V, respectfully.

The dissertation discusses the motivation behind scaling the MOSFET transistor and the challenges that scaling introduces to analog circuit design in general. The voltage reference is a basic building block for most analog circuits, and therefore must be able to scale with technology. The temperature characteristics of silicon and MOSFET devices are discussed in order to develop analog circuits that demonstrate specific and controlled temperature characteristics. By describing the temperature characteristics of these devices with polynomials, the Method of Least Squares may be used to determine gain coefficients that give an optimum temperature independent voltage reference.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 INTRODUCTION

As Systems on a Chip (SOCs) become larger and more complex, they are able to support higher performing and more complex systems. Cell phones, broadband communications, and audio-video wireless systems are just a few examples of SOC applications. SOCs that are being developed in Ultra Deep Submicron (UDSM) processes use transistors lengths less than 100 nanometers (nm). Devices of this scale inhibit performance capabilities of analog circuits that are traditionally used in larger dimensioned processes [5, 7]. As the technology roadmap continues to follow Moore's Law into UDSM processes, there is a need to develop next generation analog circuit architectures that support SOCs.

Voltage references, current references, and time references are basic building blocks for analog systems [7, 15, 33, 74]. Analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) use voltage and current references. Phase-locked loops (PLLs) use time references as well as current references. Because a current reference can be and is often generated from a voltage reference, the voltage reference is arguably the more fundamental circuit of the two. Therefore, in order for analog circuits to continue to be integrated into the next generation SOCs, a voltage reference design in UDSM process technologies must be identified.

Voltage reference circuits provide a known voltage that is temperature and power supply independent. This stable voltage is used by ADCs and DACs to produce predictable signals in both the digital and analog domains. ADCs sample analog signals and compare them to the reference voltage. The measured difference between these two values is encoded into a digital word. Similarly, a DAC will convert a digital word into

an analog signal based on a reference voltage to digital word mapping. The absolute accuracies of ADCs and DACs are dependent on the voltage reference.

As technology scales to smaller dimensions, analog circuits must adapt to electrical changes. For example, power supplies follow the scaling trend and are reduced in each new generation. In larger dimension processes, such as 0.5μ and 0.35μ processes, five volt power supplies were possible. However, as the devices scale to smaller dimensions, the maximum operating power supply has been reduced to approximately 1 volt and will soon be less than one volt.

1.2 VOLTAGE REFERENCE TECHNOLOGY

The concept of generating a voltage that is stable with variations in temperature is straightforward. A voltage or current that decreases linearly with increases in temperature is referred as a Complementary to Absolute Temperature (CTAT) signal. Correspondingly, a voltage or current that increases linearly with increases in temperature is referred as a Proportional to Absolute Temperature (PTAT) signal. By adding an appropriately scaled PTAT signal to a CTAT signal, it is possible to generate a signal in which the signal temperature differences cancel. Typically, the signal with the smaller variation with respect to temperature is scaled in amplitude so that the CTAT and PTAT temperature responses are equal.

The generation of a perfect CTAT signal or a perfect PTAT signal is challenging and invariably circuits that are used to generate such signals will demonstrate some nonlinearity with temperature. This nonlinearity introduces a residual temperature dependence when these signals are added, but this residual temperature dependence will be small if the nonlinearities are small. Unless specified to the contrary, the terms CTAT and PTAT as used herein, designate signals that are either appropriately linear with temperature or that have a small nonlinearity with temperature.

The first proposal to sum CTAT and PTAT voltages generated from bipolar devices, was published in 1964 [70]. The base to the emitter voltage of a bipolar device, V_{be}, has a CTAT response. By taking the voltage difference between two V_{be}s, ΔV_{be}, that have different current densities, a PTAT voltage can be generated. The typical V_{be} response with temperature shows a $-1.9 mV/^{o}C$ temperature coefficient and the typical ΔV_{be} response with temperature shows a $0.1 mV/^{o}C$ (device area dependent) temperature coefficient [15, 33, 74]. The standard equation for a voltage reference that uses bipolar devices is

$$V_{ref} = V_{be} + K\Delta V_{be} \text{ [15, 33, 70-75],} \tag{1.1}$$

where K is the gain needed to compensate for the difference in the magnitude of the two temperature coefficients. Figure 1.1 shows a well known circuit that can be used to sum the bipolar voltages, as shown in Equation 1.1, to produce a temperature stable voltage [33, 74]. By choosing appropriate component values and device sizes, the final

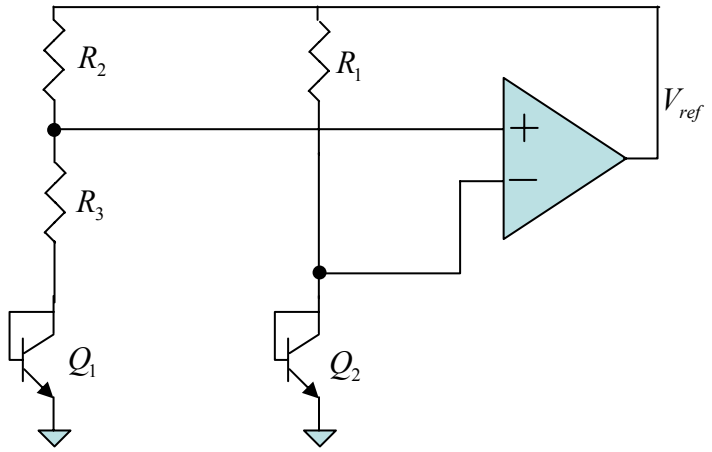

Figure 1.1: Voltage Reference: $V_{ref} = V_{be} + K\Delta V_{be}$

summation of the CTAT and scaled PTAT voltages result in V_{ref} being temperature stable at approximately 1.25 volts. The voltage reference circuit in Figure 1.1 is often

termed a bandgap reference because the resultant output voltage is dependent upon the bandgap voltage of silicon. Most existing voltage references that exploit the temperature dependent properties of a bipolar device are appropriately termed bandgap references as well. The OPAMP in Figure 1.1 must operate with a power supply voltage large enough to permit the required output level. With the trend of UDSM power supplies decreasing below one volt, this circuit is incapable of providing an output voltage of 1.25 volts.

Several architectures are proposed in the literature that are capable of producing a temperature independent reference below 1.25 volts. In fact, recent publications describe voltage references operating at one volt [76-80]. These architectures sum PTAT and CTAT currents that are converted from V_{be} and ΔV_{be} voltages. The resulting temperature independent voltage is no different than if the voltages had been summed, except that they may be scaled to any arbitrary voltage. The advantage of current summing is that the power supply voltage requirements are no longer determined by the 1.25 volts. Figure 1.2 is a good example of a low voltage bipolar voltage reference that uses current summing techniques [76].

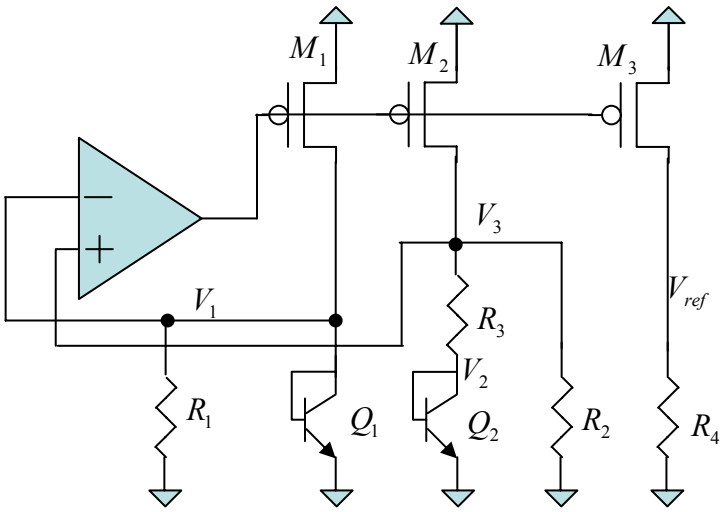

Figure 1.2: One Volt CMOS Voltage Reference

One problem with this voltage reference is the limit on the minimum power supply voltage that can be used when attempting to operate with supply voltages well below 1 volt. This limit is inherently related to the I-V characteristics of a bipolar device. Using KVL on the voltage reference shown in Figure 1.2, it is observed that the V_{be} of Q_1 plus the V_{ds} of M_2 will determine the minimum power supply requirements. The V_{be} of a bipolar device may be as high as 0.8 volts in the desired temperature range while it is difficult to achieve the minimum V_{ds} voltage of a MOSFET in saturation much below 0.1 volts. Therefore, the minimum operating voltage of this circuit, as well as any others that incorporate a bipolar device, is limited to approximately 0.9 volts. This problem creates a lower limit on the supply voltage of bipolar-based references and requires innovative circuit architectures to break through this barrier to address future technologies.

New architectures have been proposed that do not use bipolar devices [66, 81-84]. These are CMOS based circuits and they remove the limit set by the V_{be} of a bipolar device and offer the potential to scale with technology. The amount of power supply difference between V_{dd} and V_{ss} is related to the threshold voltage (V_{th}) and V_{ds} of MOSFET devices. As will be shown in Chapter 2, the threshold voltages scale with technology. However, the ratio of the power supply to threshold voltage scaling is not unity. The power supply decreases at a faster rate then the threshold voltage. This relationship is due to non-ideal effects of transistors as they are scaled to smaller dimensions.

Voltage references that have the capability to scale with technology and still provide a stable, temperature independent voltage that performs as well as or better than existing bipolar bandgap circuits are needed. In an attempt to meet this need, this work focuses on the development of CMOS voltage references that have no more than two devices stacked between the power supplies.

1.3 DISSERTATION ORGANIZATION

The focus of the dissertation is a scalable voltage reference that can be used in current and future technologies. The contributions of this dissertation are:

- A scalable CMOS voltage reference system design and methodology.
- A voltage reference that uses a maximum of two devices stacked between the power supplies. Results are shown for a voltage reference that operates from a standard power supply voltage down to just two threshold voltages of a MOSFET in 0.18μ and $90nm$ technologies.
- The temperature dependence of the CMOS voltage reference is reduced by more than 10x, resulting in performance competitive with existing bipolar voltage references.

Chapter 2 discusses transistor scaling theory. First, ideal scaling trends and the benefits offered to both digital and analog circuits are covered. Following the discussion of benefits, non-ideal effects are discussed and addressed. Within this chapter, it should become apparent that the need to reduce power consumption and chip area, while increasing the frequency at which the chip may operate, is driving transistor scaling. It should also become apparent that the architecture developed and shown in this dissertation uses at most two MOSFET devices between the power supplies to address the device threshold voltage to V_{dd} relationship.

Chapter 3 discusses the fundamental physical parameters that are temperature dependent and affect the MOSFET temperature dependent characteristics. The temperature dependence of the energy bandgap, intrinsic carrier concentration, and carrier mobility are discussed. These insights allow for the development of temperature dependent circuits that will be used to generate the scalable UDSM high performance voltage reference.

Chapter 4 covers various CTAT and PTAT circuits that may be used to generate a temperature independent reference. In this chapter, basic MOSFET equations are introduced that, to first order, allow for easy insight into a circuit's temperature characteristics. With these equations, several circuits are evaluated that meet the needs of the technology scaling trends.

Chapter 5 explains, at the system level, the proposed voltage reference architecture. A simple approach is given in which a higher-order voltage reference may be designed with the circuits described in Chapter 4. The order of the voltage reference is unlimited, and in theory could produce a perfectly temperature independent voltage reference. However, with non-ideal effects and device mismatch, it is impossible to get continued improvement. Using the theoretical temperature characteristics of parameters presented in Chapter 3 and using circuit equations presented in Chapter 4, the anticipated performance of the new architecture is discussed. This discussion shows that using this architecture offers potential to improve the temperature stability by over an order of magnitude beyond the performance reported in the literature.

Chapter 6 uses the principles and theory presented in previous chapters to complete circuit design of 1^{st}, 3^{rd}, and 5^{th} order voltage references in a 0.18μ process with a power supply of 0.9 volts. In order to show the scalability of this circuit, the same architecture is presented in a $90nm$ process with a power supply of 0.4 volts. Realistic process variations are included in the simulations. A short discussion on noise and possible methods of calibration are presented.

Chapter 7 discusses final conclusions and future work.

Chapter 2: Trends in Process Technology

2.1 INTRODUCTION

This chapter focuses on the trend of shrinking the MOSFET's physical size. Ideal transistor scaling theory and non-ideal effects of the MOSFET as a result of shrinking its geometrical parameters are discussed. First, in section 2.2, the ideal theory behind scaling the MOS device is discussed. Ideal scaling theory indicates that circuits have the ability to operate at a higher frequency while reducing the power consumption and area. This is the motivation for shrinking the MOS device. However, due to non-ideal effects, the scaled transistor does not perform consistent with the ideal scaling theory. However, the ideal scaling theory is useful as a goal for performance. Section 2.3 discusses some of the major non-ideal effects that result from shrinking the MOS device to smaller technologies. The conclusion of this chapter is that Systems On a Chip (SOC) still benefit from shrinking the MOS device, but leakage currents increase, causing unwanted power drain, and the V_{dd} to V_{th} ratio decreases. This provides the motivation to create a voltage reference circuit that uses only two stacked devices between the power supplies.

2.2 IDEAL TRANSISTOR SCALING THEORY

2.2.1 Introduction

By scaling the MOSFET's physical and geometrical properties so that all of the electric fields scale proportionally, significant circuit benefits are realized and have paved the way for complex SOCs. In addition to scaling the dimensions and doping concentrations of the device, the power supply (V_{dd}), drain to source voltage (V_{ds}), threshold voltage (V_{th}), and the gate to source voltage (V_{gs}) are scaled as well. Figure 2.1 shows the physical, geometrical, and potential voltage changes that are part of the

ideal device scaling process. Geometrical dimensions that are reduced by the reciprocal of the scaling factor α are the transistor's gate width (W), gate length (L), gate oxide thickness (t_{ox}), and the source and drain region's depth (d) and width (k). The doping concentration of the drain, source, and substrate region (N_A, N_D) of the device are increased by the scaling factor α. The control voltages of the device (V_{gs}, V_{ds}, V_{th}, and V_{dd}) are all required to scale down by α as well. With the assumptions that are summarized in Table 2.1, a simple analysis can show why the scaling of the transistor is so attractive to SOCs.

Table 2.1: Ideal Scaling Properties of MOSFET Device

Device Property	Scaling Factor
Device Width W	$\frac{1}{\alpha}$
Device Length L	$\frac{1}{\alpha}$
Gate Oxide Thickness τ_{ox}	$\frac{1}{\alpha}$
Source/Drain Region Depth d	$\frac{1}{\alpha}$
Source/Drain Region Width k	$\frac{1}{\alpha}$
Source/Drain Doping Concentration	α
Substrate/Channel Doping Concentration	α
V_{dd}	$\frac{1}{\alpha}$
V_{ds}	$\frac{1}{\alpha}$
V_{th}	$\frac{1}{\alpha}$
V_{gs}	$\frac{1}{\alpha}$

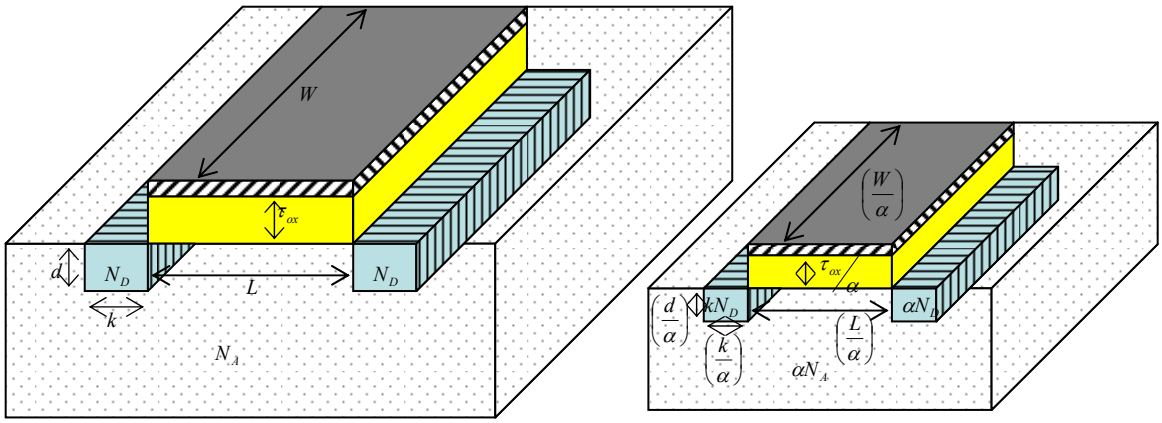

Figure 2.1: Ideal MOSFET Scaling by Factor α

2.2.2 Drain Saturation Current

The performance-to-device-scaling relationship is highly characterized by the drain current of the device operating in the saturation region and is discussed in more detail further in this section. Device characteristics in the saturation region are described by the following voltage-to-current relationship [11, 12, 15, 18]:

$$I_d = \frac{\mu_{eff} C_{ox} W_{eff}}{2L_{eff}} \left(V_{gs} - V_{th}\right)^2 \cdot f\left(V_{gs}, V_{th}, R_s, \nu_{sat}, N_A, V_{ds}, \lambda\right). \tag{2.1}$$

The drain current, I_d, of an n-channel device is related to the following process and physical parameters: Electron effective mobility μ_{eff}; Oxide capacitance per unit area C_{ox}; Effective device width W_{eff}; Effective device length L_{eff}; Voltage gate to source V_{gs}; Voltage drain to source V_{ds}; Channel length modulation coefficient λ; Threshold voltage V_{th}; electron saturation velocity ν_{sat}; and the channel, source, and drain region doping concentration N_A and N_D. Non-ideal effects are lumped into the function $f\left(V_{gs}, V_{th}, R_s, \nu_{sat}, N_A, V_{ds}, \lambda\right)$ and are assumed to be equal to one for this analysis. Gate capacitance per unit area C_{ox}, is calculated by

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}, \tag{2.2}$$

where ε_{ox} is the permittivity of the gate silicon dioxide and t_{ox} is the gate oxide thickness. Preserving the electric fields of the MOSFET device by following the scaling guide summarized in Table 2.1, the drive current may be shown to scale as follows:

$$I_{d\_scaled} = \frac{\left(\mu_{eff}\right)\left(\frac{\varepsilon_{ox}\alpha}{t_{ox}}\right)\left(\frac{W_{eff}}{\alpha}\right)}{\frac{2L_{eff}}{\alpha}}\left(\frac{V_{gs}}{\alpha} - \frac{V_{th}}{\alpha}\right)^2; \tag{2.3}$$

$$I_{d\_scaled} = \frac{1}{\alpha}\frac{\mu_{eff}C_{ox}W_{eff}}{2L_{eff}}\left(V_{gs} - V_{th}\right)^2; \tag{2.4}$$

$$I_{d\_scaled} = \frac{1}{\alpha}I_d \tag{2.5}$$

The drain saturation current (I_{dsat}) of an n-channel device can be defined as the measured drain current when the drain and gate of a device are tied to Vdd and the source of the device is grounded [13, 14]. This is a useful figure of merit (FOM) that can be used when evaluating different processes in different fabrication facilities. Again, noting the assumptions of Table 2.1 and allowing $V_{gs} = Vdd$ in equation 2.1, the saturation current, as defined above, may be shown to scale as

$$I_{dsat\_scale} = \frac{1}{\alpha}I_{dsat}. \tag{2.6}$$

2.2.3 Propagation Delay

A good figure of merit for digital logic frequency performance is the propagation delay time, τ_{pd}. The propagation delay time through a logic gate, such as an inverter, will limit the maximum frequency of an SOC. The maximum frequency capability of digital logic circuits is limited by I_{dsat}, as described above, and by capacitive loading of MOSFET devices.

A digital gate is capacitively loaded by its own drain capacitance and the gate capacitances of the following devices. A diagram of the locations of the drain capacitances and gate loading capacitances is shown in Figure 2.2. In this example, an inverter is driving an inverter as its load. As can be observed in Figure 2.2, the drain capacitances of devices M1 and M2 plus the input capacitances of M3 and M4 contribute to the total capacitance that must be driven by the first inverter.

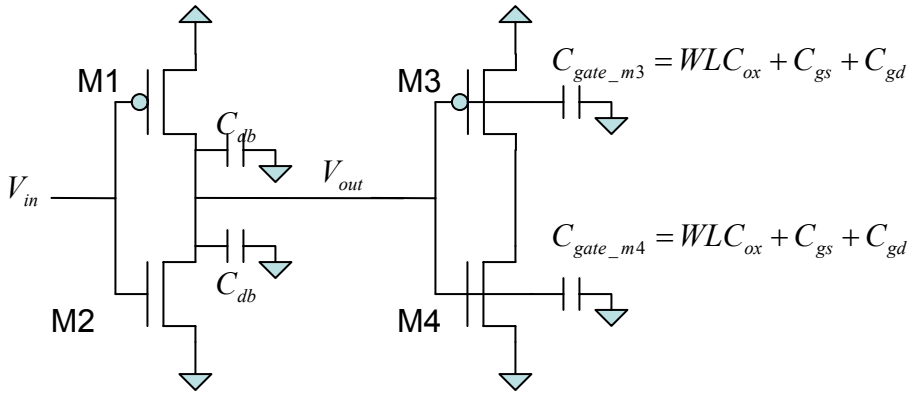

Figure 2.2: Inverter Gate Logic with Loading Capacitance

Several sources of capacitance are observed at the gate input of a MOSFET. Figure 2.3 shows the cross section of an n-channel device and where the dominant capacitors exist for a long channel device. There are three major contributors to the input capacitance of the gate. The gate-to-channel capacitance is the dominant capacitance and is calculated as

$$C_{gch} = WLC_{ox}. \tag{2.7}$$

This capacitance is described as $C_{gch} = C_{gs} + C_{gd}$. Two additional contributions to the input gate capacitance are fringing capacitances. The fringing capacitances result from the gate slightly overlapping the drain and source regions. The fringing capacitance, in the long channel device case, from the gate to source (C_{gsf}) and from the

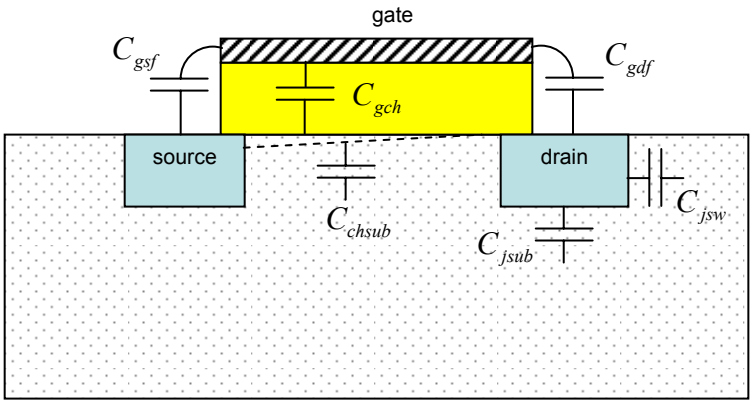

Figure 2.3: MOSFET Cross Section with Capacitors

gate to drain (C_{gdf}) are negligible in regards to their contribution to the signal propagation delay and are ignored for this analysis.

At the drain node of the device that is driving the signal, there are typically two dominant sources of capacitance. As shown in Figure 2.3, there is the capacitance from the gate to drain, C_{gdf}, and the PN junction depletion capacitances, C_{jsub} and C_{jsw}. As was the case before, C_{gdf} is negligible and is ignored. The sidewall capacitance, C_{jsw}, is calculated around the perimeter of the drain region. The junction capacitance, C_{jsub}, is capacitance that results from the PN junction that interfaces between the drain region and the substrate of the chip.

With all of the physical and geometrical properties of the MOSFET being reduced, it can be shown that the depletion capacitance will be reduced as well. In general, the depletion capacitance of a PN junction (C_j) is dependent on the surface area (A), the width of the depletion region (W_d), and the permittivity of silicon (ε_{si}). The general equation for a PN junction capacitance is calculated as [11, 12, 15]

$$C_j = \frac{\varepsilon_{si} A}{W_d}. \tag{2.8}$$

The depletion width of the device is dependent on the reverse bias of the PN junction and the doping concentration of the P and N regions. The depletion width of a PN junction is expressed by [11, 12, 15]

$$W_d = \left[\frac{2\varepsilon_{si}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right) \left(\phi_{bi} + V_R \right) \right]^{0.5} . \tag{2.9}$$

The built in potential, ϕ_{bi}, of the PN junction is due to the diffusion of the majority carriers of the PN junction. This junction potential difference is dependent on the doping concentration, temperature, and the intrinsic carrier concentration density ni. The resulting potential difference is

$$\phi_{bi} = \frac{kT}{q} \ln\left(\frac{N_A N_D}{ni^2} \right). \tag{2.10}$$

The parameter V_R is defined as the drain to bulk potential. It is assumed that the bulk and source potential are tied to the same power supply. The assumptions of $\phi_{bi} << V_R$ and $V_R = V_{ds}$ result in an approximation for W_d such that

$$W_d \approx \left[\frac{2\varepsilon_{si}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right) \left(V_{ds} \right) \right]^{0.5} . \tag{2.11}$$

Using the scaling factors outlined in Table 2.1, the scaled depletion width is calculated as

$$W_{d\_scaled} = \left[\frac{2\varepsilon_{si}}{q} \left(\frac{1}{\alpha N_A} + \frac{1}{\alpha N_D} \right) \left(\frac{V_{ds}}{\alpha} \right) \right]^{0.5} , \tag{2.12}$$

$$W_{d\_scaled} = \frac{1}{\alpha} \left[\frac{2\varepsilon_{si}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right) \left(V_{ds} \right) \right]^{0.5} , \tag{2.13}$$

$$W_{d\_scaled} = \frac{1}{\alpha} W_d . \tag{2.14}$$

The scaling effect on the total capacitance at the drain nodes of M1 and M2, as shown in Figure 2.2, may now be calculated. To simplify the calculation, all of the drain area and gate area that affects the capacitive loading is normalized. Referring to Figures

2.1 and 2.2 and expanding equations 2.7, 2.8, and 2.11, the total loading capacitance may be calculated as follows:

$$C_{load} = C_{gc} + C_{jsub} + C_{jsw}; \tag{2.15}$$

$$C_{load} = WLC_{ox} + \frac{\varepsilon_{si}(kW)}{W_d} + \frac{\varepsilon_{si}(2dk + 2dW)}{W_d}. \tag{2.16}$$

Now, using the scaling properties shown in Table 2.1 and the scaling factors of the depletion width from equation 2.14, the scaling of the capacitive load may be observed:

$$C_{load\_sclaed} = \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox}) + \varepsilon_{si}\frac{\frac{k}{\alpha}\frac{W}{\alpha}}{\frac{W_d}{\alpha}} + \varepsilon_{si}\frac{2\left(\frac{d}{\alpha}\frac{k}{\alpha} + \frac{d}{\alpha}\frac{W}{\alpha}\right)}{\frac{W_d}{\alpha}}; \tag{2.17}$$

$$C_{load\_scaled} = \frac{1}{\alpha}\left[WLC_{ox} + \frac{\varepsilon_{si}(kW)}{W_d} + \frac{\varepsilon_{si}(2dk + 2dW)}{W_d}\right]; \tag{2.18}$$

$$C_{load\_scaled} = \frac{1}{\alpha}C_{load}. \tag{2.19}$$

There are some additional capacitances that were not considered in the development of equation 2.15. The channel to bulk depletion capacitance (C_{chsub}) is ignored because it is assumed that it is charged when the channel of the device is turned on. The source depletion region capacitance is of no concern because it is already charged due to the connection to the power supply. Interconnect parasitic capacitances have been ignored. However, this is a significant issue for nanometer technologies, but this issue will be addressed later.

With all of the major contributors that affect the frequency capabilities of digital circuits of an SOC being known, the τ_{pd} scaling effect may now be investigated. From the first order relationship of

$$I = C\frac{dV}{dT}, \tag{2.20}$$

it has been shown and measured that a good approximation for the signal propagation delay time of an inverter is [13, 14, 32]

$$\tau_{pd} = \frac{C_{load} \cdot V_{dd}}{I_{dsat}}. \tag{2.21}$$

As the technology is scaled, the signal propagation delay is reduced:

$$\tau_{pd\_scaled} = \frac{\frac{C_{load}}{\alpha} \cdot \frac{V_{dd}}{\alpha}}{\frac{I_{dsat}}{\alpha}}; \tag{2.22}$$

$$\tau_{pd\_scaled} = \frac{1}{\alpha}\tau_{pd}. \tag{2.23}$$

As a result of scaling the devices down by increasing the scaling factor, the signal propagation delay is reduced, thus increasing the circuit's ability to operate at a higher frequency.

2.2.4 Digital Gate Power Density

The power consumption of a digital gate or digital circuit is dependent on the rate at which electrons are being used to charge and discharge the capacitive load of the digital circuit. Assuming that the cross-over current is negligible, it is easily shown that the power, P, consumed by a digital circuit is [13, 35]

$$P = C \cdot V_{dd}^2 f. \tag{2.24}$$

The clocking frequency (f) of the digital circuit may be assumed to be related to the inverse of the propagation delay time that was calculated in equation 2.21. Expanding equation 2.24 to include the scaling factor that was calculated above, the following power consumption to scaling relationship is observed:

$$P_{scaled} = \frac{C}{\alpha}\frac{V_{dd}^2}{\alpha^2} \cdot \alpha f; \tag{2.25}$$

$$P_{scaled} = \frac{1}{\alpha^2} C \cdot f \cdot V_{dd}^2; \tag{2.26}$$

$$P_{scaled} = \frac{1}{\alpha^2} P. \tag{2.27}$$

2.2.5 Device Intrinsic Gain

A good FOM of a MOSFET device being used in an analog circuit is its intrinsic gain. The intrinsic gain of a device (A_v) is its product of transconductance (g_m) and output impedance (r_o) [11, 12, 15, 33]:

$$A_v = g_m r_o . \tag{2.28}$$

This relationship proves to be important when evaluating the capabilities of circuits such as OPAMPs and comparing them in different process technologies.

The transconductance of a device is the amount of drain current that changes with a slight change of voltage on its gate. This is calculated around a stable operating point where V_{ds} is held constant and the device is assumed to be operating in its saturation region. Thus, by taking the derivative of equation 2.1 with respect to V_{gs} and keeping in mind all of the scaling relationships developed above, the impact on g_m due to scaling may be realized [11, 12, 15, 17]:

$$g_m = \frac{\partial I_d}{\partial V_{gs}}|_{V_{ds}=\text{constant}} ; \tag{2.29}$$

$$g_m = \frac{\mu C_{ox} W}{L}\left(V_{gs} - V_{th}\right); \tag{2.30}$$

$$g_{m\_scaled} = \mu\left(\alpha C_{ox}\right)\frac{\left(\frac{W}{\alpha}\right)}{\left(\frac{L}{\alpha}\right)}\left(\frac{V_{gs}}{\alpha} - \frac{V_{th}}{\alpha}\right); \tag{2.31}$$

$$g_{m\_scaled} = g_m . \tag{2.32}$$

The output impedance of the device is dependent on how well the channel length is not modulated by V_{ds}. The modeling equation for output impedance is related to the channel-length modulation coefficient (λ) and the drain current, as given in the following equation [15]:

$$r_o = \frac{1}{\lambda I_d} . \tag{2.33}$$

The depletion region width formed by the PN junction of the channel and drain region is a function of the drain voltage. The width of the depletion region from the drain of the device will increase as V_{ds} is increased. As the depletion width increases, it reduces the effective channel length of the device. As a result, the gate voltage does not have enough charge to overcome this depletion region and effectively cuts off the channel. This phenomenon is called pinch-off. Figure 2.4 displays this relationship, where L is the effective length of the device and ΔL is the length that the channel is reduced due to the depletion width.

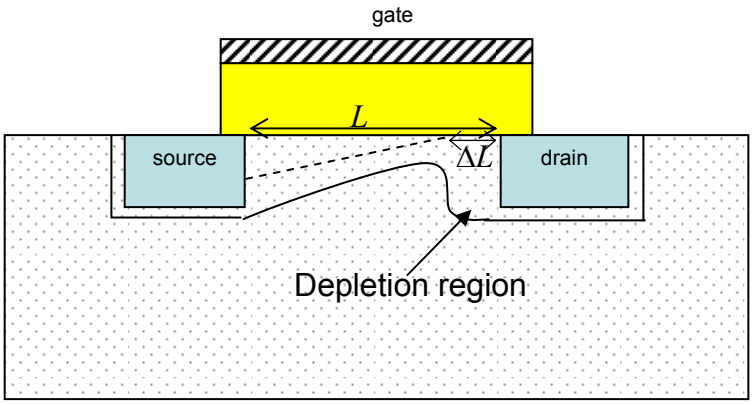

Figure 2.4: Device Showing the Depletion Region and Pinch-Off Region

The channel length modulation coefficient is used to model and describe this effect and is related with the following equation [15]:

$$\frac{\Delta L}{L} = \lambda V_{ds}. \tag{2.34}$$

As described previously, both the length of the device and the depletion region scale down linearly with the scaling factor. The drain voltage, V_{ds}, is also assumed to scale down linearly with the scaling factor. Therefore, the channel length modulation coefficient scales in the following manner:

$$\lambda_{scaled} = \left(\frac{\frac{\Delta L}{\alpha}}{\frac{L}{\alpha}} \right) \frac{1}{\left(\frac{V_{ds}}{\alpha} \right)}; \tag{2.35}$$

$$\lambda_{scaled} = \alpha\lambda \, . \tag{2.36}$$

The manner in which the output impedance of a device is affected by technology scaling may now be realized:

$$r_{o\_scaled} = \frac{1}{\alpha\lambda \frac{I_d}{\alpha}} \tag{2.37}$$

$$r_{o\_scaled} = r_o \tag{2.38}$$

From equations 2.28, 2.32, and 2.38, the intrinsic gain of a device is shown to be independent of technology scaling:

$$A_{v\_scaled} = g_{m\_scaled} r_{o\_scaled}; \tag{2.39}$$

$$A_{v\_scaled} = g_m r_o; \tag{2.40}$$

$$A_{v\_scaled} = A_v \, . \tag{2.41}$$

This is a good result for the gain of the device because in the ideal sense, there are no penalties when the devices are scaled. It would be unfavorable if the intrinsic gain of the device were to decrease as the device dimensions decreased.

2.2.6 Transit Frequency Response of Device

The transit frequency response, f_t, is a FOM for the frequency capability of a specific process for analog circuit applications. The f_t of a device is defined as the frequency at which the small signal current gain reaches unity under the conditions that the source and drain of the device are shorted to analog ground [17, 33]. Figure 2.5(a) shows the schematic diagram for this circuit and Figure 2.5(b) shows the small signal model for this calculation.

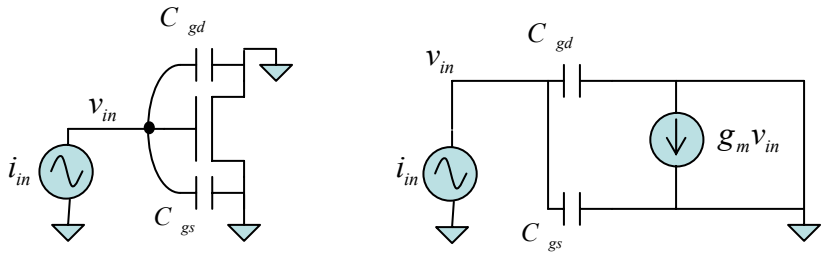

Figure 2.5: (a) Large Signal and (b) Small Signal Model to Determine f_t

In saturation, f_t is calculated with the assumption that

$$i_{in} = \sin(\omega_t t), \tag{2.42}$$

where i_{in} is the input current, $\omega_t = 2\pi f_t$, and t is time. Therefore,

$$v_{in} = \frac{-\cos(\omega_t t)}{\omega_t (C_{gs} + C_{gd})} \tag{2.43}$$

based on the relationship of equation 2.20. The total capacitance of $(C_{gs} + C_{gd})$ is approximately C_{gch} as described above in section 2.2.3. According to the small signal model and the result of v_{in}, the drain current is

$$i_d = \frac{-g_m \cos(\omega_t)}{\omega_t (C_{gch})}. \tag{2.44}$$

If the small signal current gain A_i is set equal to one and the phase relationship of the input current to the drain current is ignored, then f_t may be obtained in the following manner:

$$A_i = \frac{i_d}{i_{in}}; \tag{2.45}$$

$$1 = \frac{i_d}{i_{in}}; \tag{2.46}$$

$$f_t = \frac{g_m}{2\pi C_{gch}}. \tag{2.47}$$

The scaling of f_t may now be realized:

$$f_{t\_scaled} = \frac{g_{m\_scaled}}{2\pi C_{gch\_scaled}}; \tag{2.48}$$

$$f_{t\_scaled} = \alpha f_t . \tag{2.49}$$

The frequency response of a transistor can be expected to increase linearly with scaling. There are other methods for this type of FOM that include the gate resistance, but the end result for the ideal scaling theory remains the same [33]. Similar to the scaling results discussed above, this result has a significant impact on the SOC capabilities as the devices are scaled.

2.2.7 Ideal Transistor Scaling Theory Summary

It has been shown that by decreasing the scale of individual transistors, the power consumption and area are reduced while the frequency is increased for a specific digital circuit. Similarly, it has been shown that for such transistor scaling, the analog characteristics are either not altered or improved as well. Table 2.2 summarizes the major FOM for scaling the technologies.

Table 2.2: FOM of Ideal Scaling Properties of MOSFET Device

Device property	Scaling Factor
Digital Chip Area	$\frac{1}{\alpha^2}$
Digital Operational Frequency $\left(\frac{1}{\tau_{pd}}\right)$	α
Digital Power Consumption (assuming operation @ maximum frequency)	$\frac{1}{\alpha^2}$
Intrinsic Device Gain (A_v)	1
Intrinsic Frequency Response (f_t)	α

Now, it is obvious that scaling MOSFET devices has many advantages, including being an effective tool for increasing the complexity of systems. In the case of digital circuitry, for example, the millions of instructions per second (MIPS) of a digital signal processing (DSP) core could increase significantly by adding additional functional blocks that allow for more parallel processing. In addition, these SOCs may run at faster speeds with the sacrifice of increased power consumption.

Analog circuitry, in the meantime, is capable of maintaining the same performance specifications while increasing its bandwidth. This is an ideal situation for hand held portable device applications that take advantage of ideal scaling. Thus, with the ability to operate faster while maintaining or improving circuit performance, it becomes possible to have longer lasting cell phone calls with broadband applications such as internet access. Unfortunately, though, the physics of the device changes substantially in submicron dimensions. This results in some challenges to keeping the processing trends consistent with the ideal theory of device scaling.

2.3 NON-IDEAL TRANSISTOR SCALING EFFECTS

2.3.1 Short Channel Effects

As the geometries of the MOSFET devices are reduced, physical problems become apparent due to short channel effects. As the device geometry is scaled, the depletion widths of the source and drain regions become a significant percentage of the channel. The current flow through the device is controlled by both the electric field from the gate and the electric field from the drain to the source. The modulation of the channel length is substantial and the electric field from the gate is required to compensate for this problem. Other significant effects involve the source and drain dimensions. These dimensions cannot scale linearly with the device due to an unacceptable increase in

source and drain resistance. The built in potential, ϕ_{bi}, is not negligible, as was assumed in the development of equation 2.11. Although scaling the technology is difficult due to the non-ideal effects of the short channel devices, the results are still positive.

Several relationships were assumed in order to complete some of the evaluations in the following sections. Table 2.3 shows a table of the relationships that are used when data is presented in the following sections. By using the standard processing equations and these relationships, a general non-ideal trend may be presented that helps illustrate the severity of the short channel effects [9, 12, 15]. Note that the relationships in Table 2.3 are approximate and are being used to demonstrate a specific problem with scaling the transistor.

Table 2.3: Estimated Physical Relationships

Parameter relationship	relationship
τ_{ox} as a function of L_{eff}	$t_{ox} = \frac{L_{eff}}{45} \cdot 10^{10}\, angstroms$
Destructive Oxide Breakdown (dob)	$dob \approx 0.07 \frac{Volts}{angstroms}$
Power Supply Limit (psl) (to avoid dob)	$psl \le \frac{2}{3} V_{dd}$
Threshold to Power Supply	$V_{th} = \frac{1}{5} V_{dd}$

2.3.2 Threshold Voltage

The threshold voltage ideally scales with the technology, as described in chapter 2.2. However, the scaling of the threshold voltage introduces serious problems due to the drain leakage current when the device is turned off. Digital circuit blocks are

generally powered down when they are not operational. In general, powering off digital circuits is effectuated by turning off clocks to the digital blocks. This forces at least one device in every logic gate between V_{dd} and V_{ss} to have its V_{gs} voltage equal to zero volts. Therefore, there is no current due to charging and discharging the digital gate's capacitive loads and there is no low resistance path between the power supplies.

The "off" current of a device (I_{off}) is the drain current when $V_{gs} = 0$. I_{off} is dependent on V_{th}, N_A, N_D, the PN junction diode area of the drain region (W, d, k), t_{ox}, and V_{ds}. Some of the "off" current is attributed to diffusion current across the PN junction of the drain and bulk region and some is attributed to the electrons punching through the channel as the source and drain regions get closer. Additionally, the tunneling current through the gate oxide from the drain to the gate increases as t_{ox} is reduced. However, the most significant contribution to I_{off}, is the subthreshold current.

A device is operating in the subthreshold region when $|V_{gs}| \le |V_{th}|$. As a result of the polysilicon gate and doped channel work function difference, minority carriers exist at the oxide interface creating a depletion region. With the existing minority carriers at the gate oxide interface and an electric field imposed by V_{ds}, the drain current has an exponential dependence on V_{gs}, V_{th}, and V_{ds}. The subthreshold current is modeled by

$$I_d = \frac{\mu_{eff} C_{dep} W_{eff} \phi_t^{\,2}}{L_{eff}} \exp\left(\frac{V_{gs} - V_{th}}{\phi_t \left(1 + \frac{C_{dep}}{C_{ox}} \right)} \right) \left(1 - \exp\left(\frac{-V_{ds}}{\phi_t} \right) \right), \tag{2.50}$$

where ϕ_t is the thermal voltage and C_{dep} is the capacitance from the channel depletion region to substrate junction [10, 11, 12, 13, 15]. The thermal voltage is the well known relationship

$$\phi_t = \frac{kT}{q}, \tag{2.51}$$

where k is Boltzmann's constant $\left(8.62 \cdot 10^{-5}\, eV/K\right)$, T is the temperature in Kelvin, and q is the magnitude of electron charge $\left(1.6 \cdot 10^{-19}\, J\right)$. The depletion capacitance is approximated by [56]

$$C_{dep} = \frac{\sqrt{\varepsilon_{si} q N_A}}{\sqrt{4\phi_{bi\_ch}}}, \tag{2.52}$$

where ε_{si} is the permittivity of silicon and ϕ_{bi_ch} is the built in potential from the depletion region of the channel to the substrate junction.

To show the scaling impact of V_{th} on I_{off}, three plots are shown in Figure 2.6. The three plots are generated by scaling all or some of the properties relative to the parameters set by the 0.18μ technology. The I_{off} current for the 0.18μ process is used as the reference point to show the concerns with scaling V_{th}. To analytically observe the effects of scaling I_{off}, it is assumed that $V_{ds} = V_{dd}$ and $V_{gs} = 0$ so that:

$$I_{off\_scaled} = \frac{\mu_{eff}\left(\sqrt{\alpha} C_{dep}\right)\frac{W_{eff}}{\alpha}\phi_t^{\,2}}{\frac{L_{eff}}{\alpha}} \exp\left(\frac{-\frac{V_{th}}{\alpha}}{\phi_t\left(1+\frac{\left(\sqrt{\alpha} C_{dep}\right)}{\alpha C_{ox}}\right)}\right)\left(1-\exp\left(-\frac{\frac{Vdd}{\alpha}}{\phi_t}\right)\right); \tag{2.53}$$

$$I_{off\_scaled} = \frac{\mu_{eff}\left(\sqrt{\alpha} C_{dep}\right) W_{eff}\phi_t^{\,2}}{L_{eff}} \exp\left(\frac{-\frac{V_{th}}{\alpha}}{\phi_t\left(1+\frac{C_{dep}}{\sqrt{\alpha} C_{ox}}\right)}\right)\left(1-\exp\left(-\frac{\frac{Vdd}{\alpha}}{\phi_t}\right)\right). \tag{2.54}$$

The scaling factor may be calculated by taking the ratio of the technology of interest relative to the 0.18μ dimension. For example, the scaling factor in Figure 2.6 when comparing the $65nm$ process to the 0.18μ process is calculated as follows:

$$\alpha = \frac{180nm}{65nm}; \tag{2.55}$$

$$\alpha \approx 2.77 . \tag{2.56}$$

To demonstrate the dominance that V_{th} has on the "off" current, a curve was plotted in Figure 2.6 that varied all of the scalable process parameters except for V_{th}. As a result, the "off" current only doubled over the generations of process technology that are being shown. However, if only V_{th} is scaled, then almost six orders of magnitude change is observed. When all of the parameters are included for observing the

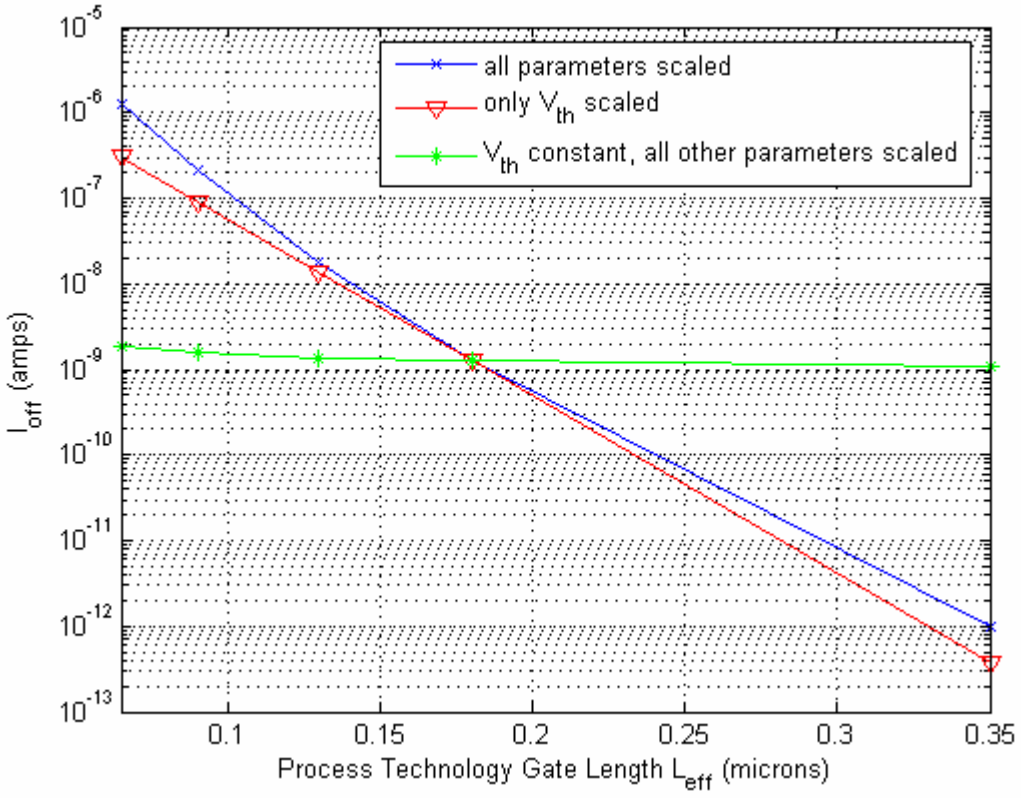

Figure 2.6: I_{off} as a Function of Technology Scaling

scaling effect, the smaller contributions from everything else is observed. Using a rough estimate of the "off" current in a larger gate length process as a reference and then showing the relative increase of current as the device is scaled down is a good example to help clarify the significance of the problem. Let's assume that there are two million

transistors on an example chip configured into a long inverter chain. By holding the input to the inverter chain at V_{dd} or V_{ss}, half of the transistors will be "on" and half of them will be "off." The current consumed by this chip in the 0.18um process would be on the order of 1.25mA. However, if this chip is scaled down to the 65nm process, the current consumption is on the order of 1 Amp. This general trend is not acceptable for low power applications and V_{th} will not be able to scale in the ideal sense.

To tackle the problem of power consumption due to the drain leakage currents, various methods are being investigated [5, 9, 10, 21, 34, 35]. One method is to have dual V_{th} voltages available in the process. The high performance circuits that require the maximum I_{dsat} can use the lower V_{th}. The slower circuits can use the higher V_{th}. With the combinational logic using the two types of devices, it is the goal to achieve an acceptable average "off" state power consumption value. Another approach is to use Dynamic Threshold-Voltage MOSFETs (DTMOS). In this case, the device has a large V_{th} when the device is "off" but has a small V_{th} when the device is "on." There are other methodologies being investigated, which are not within the scope of this section.

The drive current capability of the device is highly dependent on the threshold voltage and its ability to scale with the process. It is obvious that the threshold voltage is unable to complete this task according to the ideal scaling theory for transistors in the nanometer range. The combination of this problem with other electric fields not scaling cause other significant non-ideality problems, especially in the drain leakage through the gate oxide.

2.3.3 Oxide Thickness

The oxide thickness is reduced to produce a stronger vertical electric field in the channel. This stronger electric field gains back some gate control of the device that is lost due to the horizontal electric field getting stronger between the source and drain

regions. The horizontal electric field is increasing because the gate length is being scaled down at a faster rate than V_{ds}. Reducing the gate oxide thickness has worked well over the past generations where the depletion widths of the source and drain were not a significant portion of the channel. Unfortunately, all of the parameters of the MOSFET device do not all scale to the ideal case that is summarized in Table 2.1 [3, 4, 10, 21, 22, 32].

A closer look at the channel depletion region characteristics reveals that the inversion layer is not infinitely thin and not located exactly at the insulator-channel interface. The inversion region peaks at a distance d_{inv} from that interface [12, 47]. This results in an effective oxide thickness that is larger then the assumed t_{ox}. The electrical oxide thickness may be calculated as [5, 12, 39]

$$\hat{t}_{ox} = t_{ox} + \frac{\varepsilon_{ox}}{\varepsilon_{si}} d_{inv}. \tag{2.57}$$

The depth of the inversion layer may be approximated by

$$d_{inv} = \beta\left(Q_b + \frac{11}{32} Q_i \right), \tag{2.58}$$

where $\beta \approx 10^{-9}$, and Q_b and Q_i are the bulk and inversion layer charges per unit area, respectively, in the depletion region of the channel. The bulk and inversion layer charges are a function of the doping concentration and the surface potential, ϕ_s, that is present at the gate oxide and silicon interface. Therefore, assuming that the relationships in Table 2.1 do not hold, the effective oxide thickness is expected to increase relative to the physical thickness of the insulator. As a result, the oxide thickness will be required to be scaled even further, resulting in a relative increase of gate current leakage.

The current through the gate oxide is a result of electrons and holes tunneling through the insulator. Fowler-Nordheim (FN) tunneling and direct tunneling are the dominant sources of gate oxide current leakage [4, 21]. In both types of tunneling, the

current has an exponential dependence on the oxide thickness and electric field across the gate oxide. The International Technology Roadmap for Semiconductors (ITRS) shows the general expected trend of $\hat{t}_{ox}$ and the gate leakage current $i_{g\_leakage}$.

Effort is being made to fix the gate oxide leakage current. Alternative materials that have a high dielectric constant are being investigated. An insulator with a high dielectric constant could be made thicker while maintaining an equivalent electric field of a thinner SiO_2 gate. This would help to alleviate the gate leakage and maintain control of the transistor's channel.

2.3.4 Carrier Mobility

The carrier mobility directly impacts the drain current of a MOSFET device in both the saturation region (equation 2.1) and the subthreshold region (equation 2.50). As a carrier is accelerated in an electric field, it will at some point collide with either the lattice or a defect in the lattice [11, 12, 40, 41]. This phenomenon is generally referred to as "scattering." A crystal lattice vibrates at a frequency that is dependent on temperature. The scattering of an electron due to lattice collisions is called phonon scattering. Impurity scattering is a result of the electron colliding with a lattice defect, such as an impurity atom. The mobility of a carrier decreases as the number of collisions increases. There is a more detailed discussion of mobility in Chapter 3, but mobility needs to be addressed as a factor in the non-ideal properties of the scaling theory.

The non-ideal effects of scaling transistors result in the electric field at the surface of the channel to increase as t_{ox} decreases with technology [3, 4, 10, 21, 22]. This is partially a result of V_{dd} and V_{th} not scaling in order to maintain the expectations of scaling I_{dsat}. The increase of vertical electric field will pull the channel closer to the channel-gate oxide junction. The silicon side of the oxide-silicon interface is referred to as the "surface." At the surface, the numbers of lattice defects are larger due to the

manufacturing process of joining these two materials. Since the channel resides in an area with more defects, the effective mobility of the carriers is reduced. This non-ideal effect can be modeled by the empirical formula [12, 15, 33]

$$\mu_{eff} = \frac{\mu_o}{1 + \theta\left(V_{gs} - V_{th}\right)}, \tag{2.59}$$

where μ_o is the mobility in the absence of a gate induced electric field and θ is a fitting parameter called the mobility reduction coefficient. An approximation for the mobility reduction coefficient is [12, 15, 33]

$$\theta \approx \frac{2 \cdot 10^{-3}}{t_{ox}}\left(\frac{\mu}{volt}\right). \tag{2.60}$$

With the electric fields not scaling according to technology, the drive current capability of the MOSFET is again reduced. This will require other parameters of the MOSFET to not change in accordance with the technology, such as the V_{ds}, V_{th}, and V_{dd}, in an attempt to maintain I_{dsat} on its ideal scaling trend.

2.3.5 VDD

The reduction of the power supply voltage according to technology scaling will cause the analog power consumption to increase dramatically. The headroom voltage, V_{hr}, for analog systems can be defined as the maximum peak-to-peak voltage at which an analog signal may operate and still meet the specifications of the system. The headroom voltage can be defined as

$$V_{hr} = V_{dd} - V_{th}. \tag{2.61}$$

This definition is an over simplification because several other voltages, such as V_{ds}, have been ignored. The peak-to-peak voltage is heavily dependent on the circuit architecture, but this definition is optimistic. Also, this simplification does not detract from the conclusion of this section. Below, in Figure 2.7, a plot of V_{dd}, V_{th}, and V_{hr} is plotted to illustrate how the three voltages are related using the ideal scaling properties of Table 2.1.

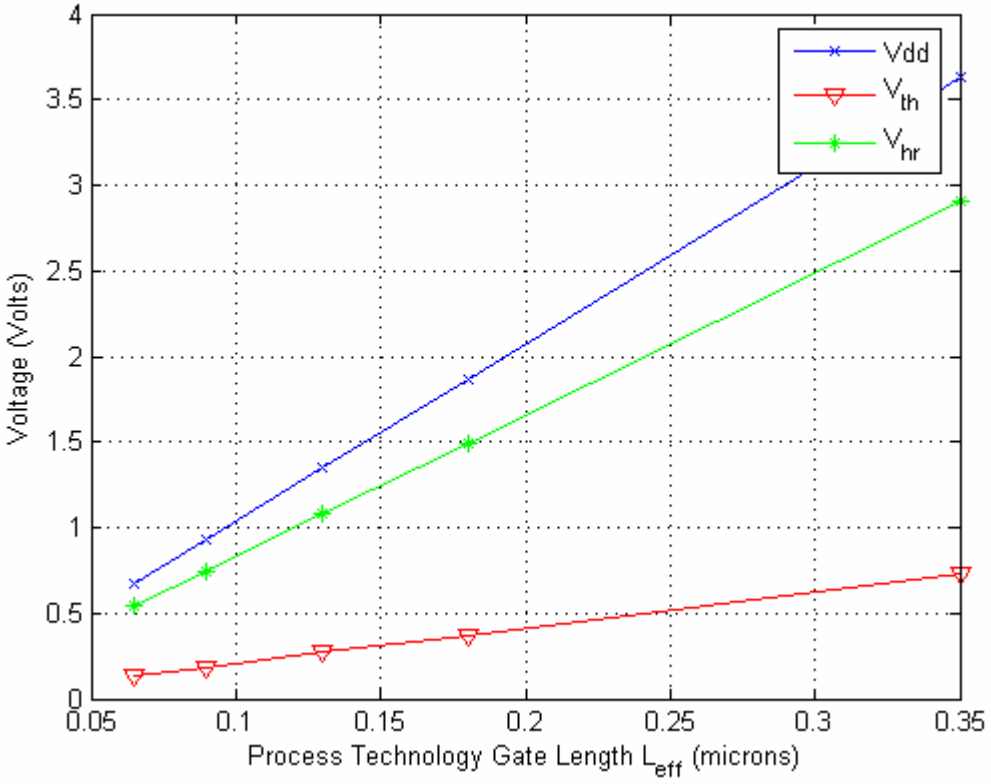

Figure 2.7: V_{dd}, V_{th}, and V_{hr} as a Function of MOSFET Process Technology

Many analog circuits stack multiple devices between V_{dd} and V_{ss} in the saturation region to enhance their gain or help suppress the power supply noise. In this example, the standard architectures for fundamental blocks are still acceptable because both V_{dd} and V_{th} are being scaled by the same factor. A serious problem that results from the headroom scaling is maintaining the Dynamic Range (DR) capabilities of the analog circuit.

Dynamic range of an analog circuit is defined as the ratio of the maximum signal power to noise power. There are several ways to define the maximum signal power. Two examples of such definitions are: 1) The 1-dB compression point (P_{1db}) of the system (mainly used in RF systems); and 2) The point at which the input power to the

analog system causes the SNR performance to fall 3dB from its peak SNR performance (Sigma-Delta ADCs). The particular applications and absolute specifications are not important in understanding the impact that the loss of headroom will cause in all of these analog systems. Assuming that power is referenced to a one ohm resistor, maximum signal power, P_{signal}, can be related to the headroom voltage as

$$P_{signal} = V_{hr}^2 . \tag{2.62}$$

It is easily shown that the maximum signal power scales down with the scaling factor as

$$P_{signal\_scaled} = \frac{1}{\alpha^2} P_{signal} \tag{2.63}$$

For simplicity, it is assumed that the only noise source is the thermal noise contribution of a single transistor. The thermal noise power of a transistor is calculated as

$$P_{\text{thermalnoise}} = v_n^2 , \tag{2.64}$$

where the thermal noise voltage is related to the transconductance of the device when it is in saturation and is expressed as

$$v_n^2 = \frac{4kT\gamma}{g_m} , \tag{2.65}$$

where γ is the noise factor [24-31]. It was shown in section 2.2.5 that the transconductance of a device does not scale according to technology. If it is assumed that the noise factor does not scale according to technology, then the thermal noise is shown to not scale according to technology:

$$v_{n\_scaled}^2 = v_n^2 . \tag{2.66}$$

The scaling relationship of the signal and noise powers has been established. In order to examine the impact that device scaling has on power consumption of analog circuits while maintaining a specified dynamic range, examine the following equation:

$$DR_{db} = 10\log_{10}\left(\frac{\frac{V_{th}^2}{4kT\gamma}}{g_m}\right). \tag{2.67}$$

Using the ideal scaling relationships shown in equations 2.59 and 2.61, the dynamic range is shown to scale as

$$DR_{db\_scaled} = 10\log_{10}\left(\frac{\frac{\frac{V_{th}^2}{\alpha^2}}{4kT\gamma}}{g_m}\right), \tag{2.68}$$

$$DR_{db\_scaled} = DR_{db} - 10\log_{10}\left(\alpha^2\right). \tag{2.69}$$

This shows that for every time the technology is scaled by ½ and the thermal noise is not compensated, the analog circuit loses 6dB of DR (i.e. 6dB per octave). The way to recover the dynamic range is to increase the current in the device to drive down its thermal noise contributions.

By solving equation 2.1 for $\left(V_{gs} - V_{th}\right)$ and substituting that into equation 2.31, it is found that

$$g_m = \sqrt{\frac{2\mu_{eff}C_{ox}WI_d}{L}}. \tag{2.70}$$

Assuming that I_d may be scaled with the technology to maintain the same dynamic range, the scaling requirement for I_d may be found with the following substitutions:

$$DR_{db\_scaled}\left(I_{d\_scaled}\right) = DR_{db}\left(I_d\right), \tag{2.71}$$

$$\frac{V_{th}^2 g_{m\_scaled}}{\alpha^2 4kT\lambda} = \frac{V_{th}^2 g_m}{4kT\lambda}, \tag{2.72}$$

$$I_{d\_scaled} = \alpha^4 I_d. \tag{2.73}$$

This is a serious problem for power consumption of analog devices, especially in portable electronics. In order to maintain the DR of a particular system the current must increase by a factor of 9dB per octave, as shown by the following relationships:

$$P_{ana\_scaled} = Vdd_{scaled} \cdot I_{scaled}; \tag{2.74}$$

$$P_{ana\_scaled\_db} = 10\log_{10}\left(\frac{Vdd}{\alpha}\alpha^4 I_d\right); \tag{2.75}$$

$$P_{ana\_scaled\_db} = P_{ana} + 10\log_{10}\left(\alpha^3\right). \tag{2.76}$$

To demonstrate the power consumption problem, an analog system is assumed that has an arbitrary DR requirement. The power requirement to meet the specified DR is 1mA for a 0.35um process. Using the relationships above, Figure 2.8 was generated to illustrate the equivalent power consumption for that particular analog circuit as it is scaled from the 0.35um process to 65nm.

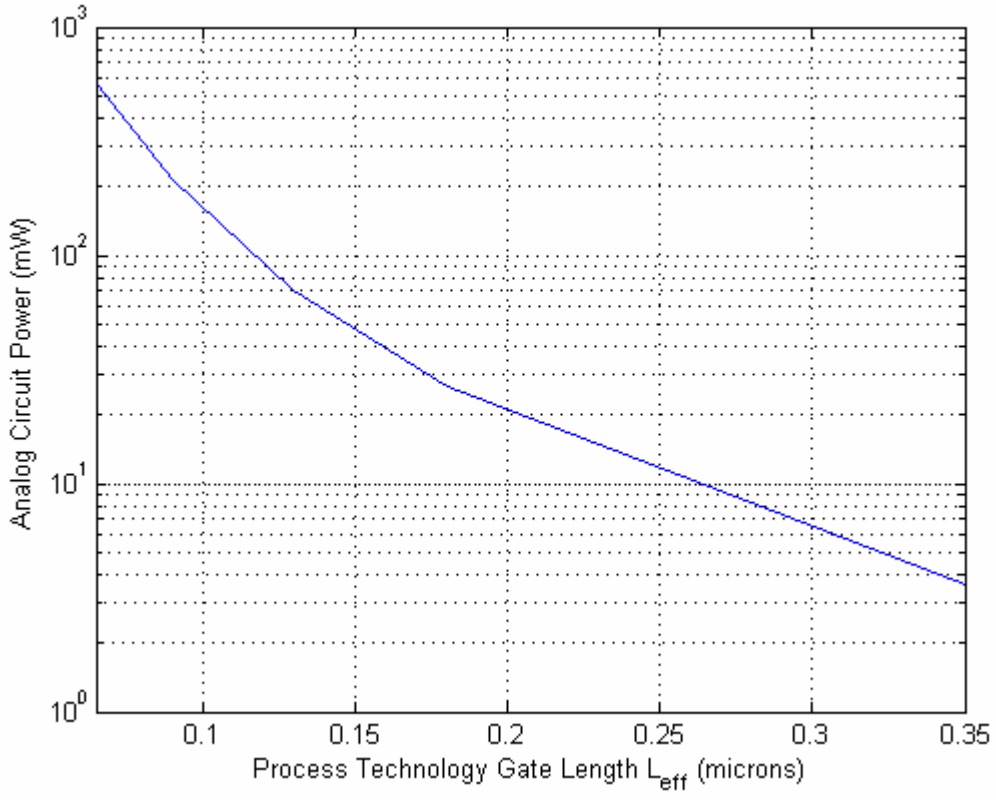

Figure 2.8: Analog Power Consumption to Maintain DR Throughout Scaling

Observing the log scale in Figure 2.8, the power consumption of this analog circuit was originally operating at an acceptable power consumption level of 3.6mW. By scaling the circuit was down to a 65nm process, it consumes approximately 550mW of power.

With optimistic assumptions about the headroom and thermal noise of an analog circuit, it was shown that the power consumption increases significantly. The noise factor γ increases as the MOSFET shrinks [24, 86, 87, 88]. This is due to many factors, but is highly related to the decrease of the carrier mobility. Factors such as these will only add to the power consumption problem that is involved with scaling the MOSFET technology.

Another non-ideality that affects analog circuit design is the fact that V_{th} is not scaling as fast as V_{dd}. The definition used for headroom in equation 2.61 may be used to discuss this severe problem. From the simple definition of headroom and the consequences of the relationship between V_{th} and V_{dd}, analog circuits are running out of headroom as the process dimensions are shrinking. Not only does this make the DR issue presented above an optimistic view, but it forces CMOS analog circuit architectures to compensate for the reduction of headroom.

2.4 TRANSISTOR SCALING SUMMARY

The benefits of scaling the transistors in order to enhance the capabilities of SOCs become apparent with the presentation of the previous sections. As circuit area decreases and operating frequencies increase, power consumption decreases. This leaves the option to expand the functions of the SOC at the price of additional current, which is still less than an SOC that exists in a larger gate process. Analog circuit bandwidths are observed to expand, enabling SOC systems to increase their communication data

throughput. In fact, some SOCs have the ability to sample intermediate radio frequencies (IF).

As with most engineering systems, tradeoffs between non-ideal effects and performance must be made. The major focus for scaling the devices is to maintain the scaling trend for digital circuits. This results in a focus to maintain the I_{dsat} and capacitive load reduction. Unfortunately this focus has adverse effects on analog circuits. As pointed out in the previous section, the circuits are losing headroom. As the analog circuit loses headroom, it must consume more power to gain back its DR. Also, to maintain the performance, new architectures must be provided. With the loss of headroom and the requirement to maintain performance, additional area will be consumed.

Chapter 3: Temperature Dependent Devices

3.1 INTRODUCTION

In this chapter the temperature dependent parameters of CMOS devices are investigated. Theoretical equations are presented and in some cases compared to the results from the Spice simulator. The main Spice simulator used is Silvaco's SmartSpice. In all of the following discussions, it is assumed that an n-channel CMOS device is being evaluated, unless otherwise specified.

3.2 ENERGY GAP OF SILICON

The temperature dependence of the energy gap, $E_g(T)$, affects the intrinsic carrier concentration of silicon and ultimately the $I_{ds} - V_{gs}$ relationship. This temperature dependence is a result of the lattice expanding and contracting in response to the temperature [43, 44].

There exists substantial theory and experimental data on the energy bands for semiconductor materials. In general, the theory seems to fall short of explaining the experimental data. The combination of matching the theory to the experimental data has brought insight to energy band diagrams of semiconductors [43-48]. The most reliable method to model the temperature dependence of the bandgap is to use experimental data. The first proposed curve fit to the experimental data utilized the following relationship:

$$E_g(T) = E_{go} - \frac{\alpha T^2}{T + \beta} \text{ [43].} \tag{3.1}$$

Due to the margin for error and interpretation of the experimental data, there are two favored sets of parameters that are used in equation 3.1. These parameters are summarized in Table 3.1.

Table 3.1: Energy Gap Variables and References

E_{go}	α	β	reference
1.16	$7.021 \cdot 10^{-4}$	1108	[43, 50]
1.17	$4.73 \cdot 10^{-4}$	636	[11, 45]

Alternative equations have been proposed, but in all of these cases the results are fairly accurate [46]. The Spice model uses the oldest proposed set of parameters. These are $E_{go} = 1.16$, $\alpha = 7.021 \cdot 10^{-4}$, and $\beta = 1108$. The energy gap temperature dependence using these parameter values is shown below in Figure 3.1.

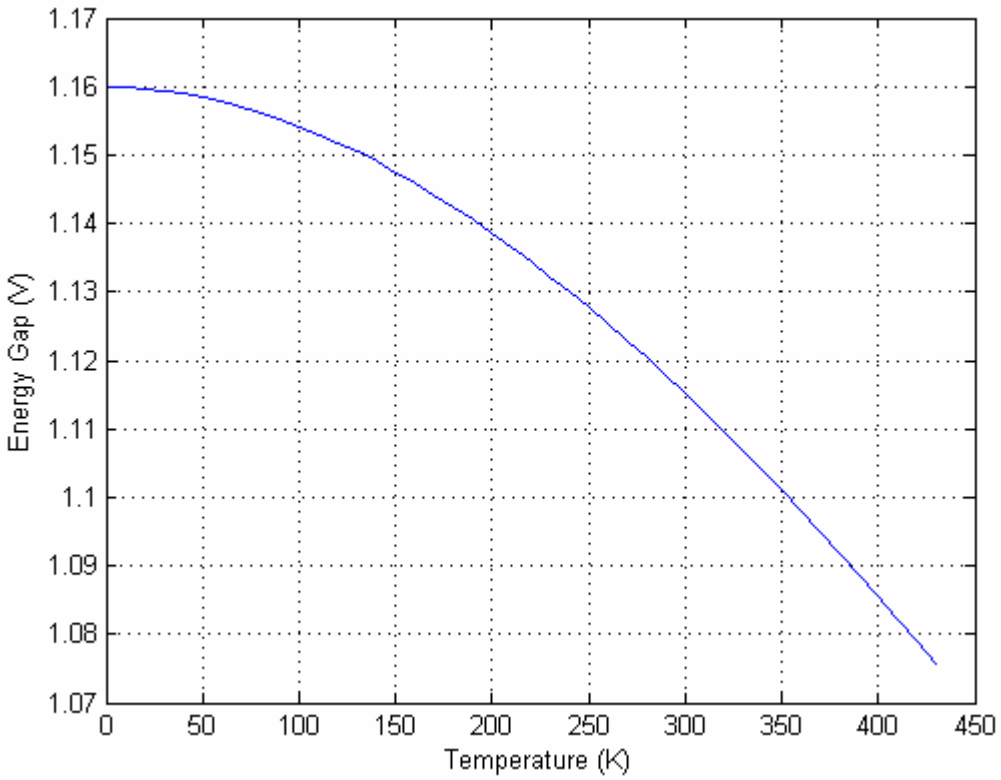

Figure 3.1: Energy Gap as a Function of Temperature

3.3 INTRINSIC CARRIER CONCENTRATION

The temperature dependence of the intrinsic carrier concentration, $n_i(T)$, will effect the Fermi potential, ϕ_f, and the threshold voltage of a MOSFET. The carrier concentration in an intrinsic semiconductor can be derived using quantum mechanics and Fermi-Dirac statistics [11, 41, 49]. The carrier concentration as function of temperature may be described by

$$n_i{}^2(T) = 4\left(\frac{2\pi kT}{h^2}\right)^3 \left(m_n^* m_p^*\right)^{\frac{3}{2}} \exp\left(\frac{-E_g(T)}{kT}\right), \tag{3.2}$$

where k is Boltzmann's constant, h is Plank's constant $\left(6.63 \cdot 10^{-34}\,\text{Joules} \cdot \text{second}\right)$, and m_n^* and m_p^* are the effective mass of electrons and holes, respectfully. The intrinsic carrier concentration is dominated by the temperature term in the denominator of the exponential. The carrier concentration of silicon, with respect to temperature is shown in Figure 3.2.

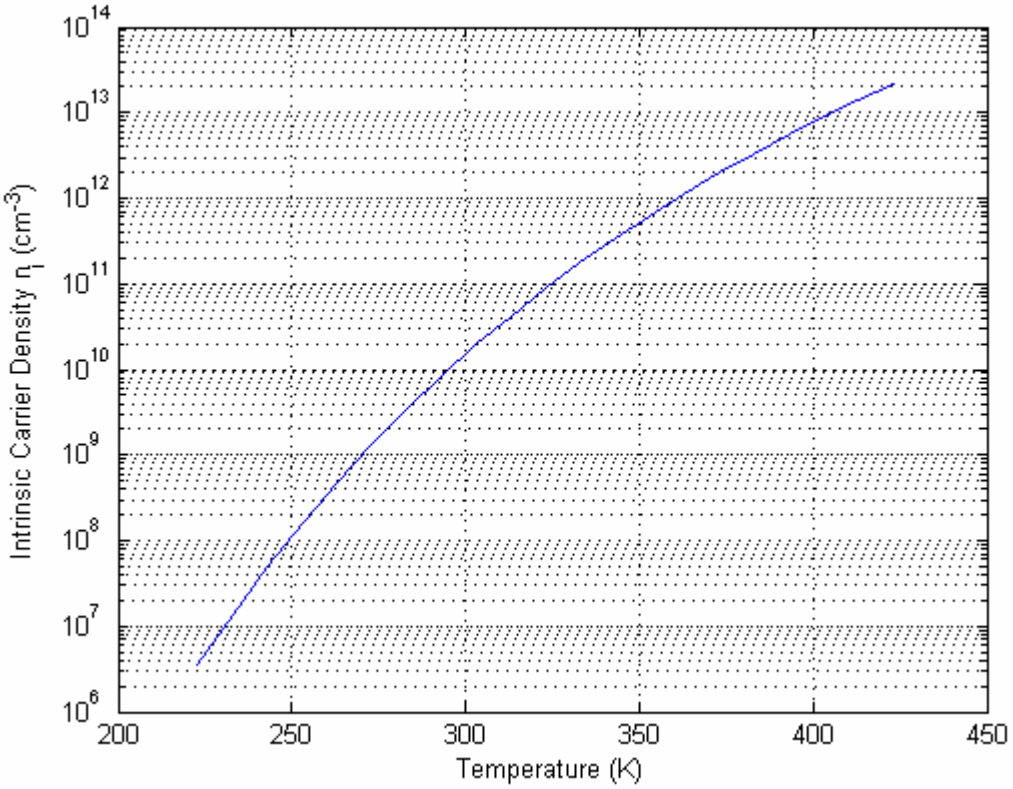

Figure 3.2: Intrinsic Carrier Density of Silicon with Respect to Temperature

3.4 THRESHOLD VOLTAGE

The threshold voltage is used in many MOSFET models to relate the drain current to the source, drain, gate, and bulk potentials. By definition, the threshold voltage is the V_{gs} that makes the surface potential, ϕ_s, approximately twice the Fermi potential, ϕ_f [12]. The threshold voltage is calculated by the following relationship,

$$V_{th} = V_{FB} + 2\phi_f + \gamma\sqrt{2\phi_f}\,, \tag{3.4}$$

where V_{th} is the threshold voltage, V_{FB} is the flatband voltage, ϕ_f is the Fermi potential, and γ is the body effect coefficient [11, 12]. The flatband voltage is expressed by

$$V_{FB} = \phi_{ms} - \frac{Q_{ox}}{C_{ox}}, \tag{3.5}$$

where ϕ_{ms} is the metal to semiconductor work function and Q_{ox} is the trapped/mobile ions at the $SiO_2 - Si$ interface. To simplify the analysis and focus on the temperature response, it is assumed that Q_{ox} is equal to zero. The gate of a MOSFET device is no longer metal but is made up of heavily doped silicon, usually called poly-silicon. Therefore, the ϕ_{ms} potential may be fully characterized by the difference between the channel and the gate Fermi potentials, so that

$$\phi_{ms} = -\frac{kT}{q}\ln\left(\frac{N_{Dgate}N_A}{n_i^{\,2}}\right). \tag{3.6}$$

The temperature independent body effect coefficient is given by the relationship

$$\gamma \equiv \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}}. \tag{3.7}$$

The doping concentration N_A is assumed to be temperature independent due to all of the electrons being ionized when the temperature is greater then $50^o K$.

The threshold voltage may now be expanded into an equation that shows all of its temperature dependence:

$$V_{th} = -\frac{kT}{q}\ln\left(\frac{N_{Dgate}N_A}{n_i^{\,2}}\right) + 2\frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) + \gamma\sqrt{2\frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right)}. \tag{3.8}$$

If this relationship is expanded even further with a defined temperature independent parameter

$$K_{ni}^{\,2} = 4\left(\frac{2\pi k}{h^2}\right)^3\left(m_n^* m_p^*\right)^{\frac{3}{2}}, \tag{3.9}$$

then

$$V_{th} = \frac{kT}{q}\ln\left(\frac{N_A}{N_{Dgate}}\right) + \gamma\sqrt{\frac{kT}{q}\left(2\ln\left(\frac{N_A}{K_{ni}}\right) - 3\ln(T) + E_{go} - \frac{\alpha T^2}{T+\beta}\right)} \tag{3.10}$$

This results in a complex relationship to temperature and is not easily reduced into a simplified equation. Unfortunately all of the terms in this expression are fairly close in value to each other. Trying to simplify this equation introduces errors that are on

the order of 50% over the temperature range. Using a higher level simulation tool, such as Matlab, a polynomial curve fit may be realized. As it turns out, the temperature dependence is very linear to a first order approximation. Spice simulators take advantage of this and actually use a first order linear approximation polynomial. The first order expression is

$$V_{th}(T) = V_{th\_tnom} + K_{vth}(T - T_{nom}), \tag{3.11}$$

where T_{nom} is the normalized temperature (such as room temperature), V_{th_tnom} is the threshold voltage at the normalized temperature, and K_{vth} is the first order temperature coefficient. The threshold temperature coefficient is negative and may be estimated to be equal to approximately $-0.8mV$ per ^{o}C, but is actually process dependent. A plot of the threshold temperature dependence from equation 3.8 and from a simulation in Spice is shown in Figure 3.3. The following parameters are used to generate this plot: $N_A = 7.38 \cdot 10^{17}/cm^3$; $N_{Dgate} = 4 \cdot 10^{19}/cm^3$; $V_{th\_tnom} = 0.46V$; and $K_{vth} = -0.224V/^{o}C$.

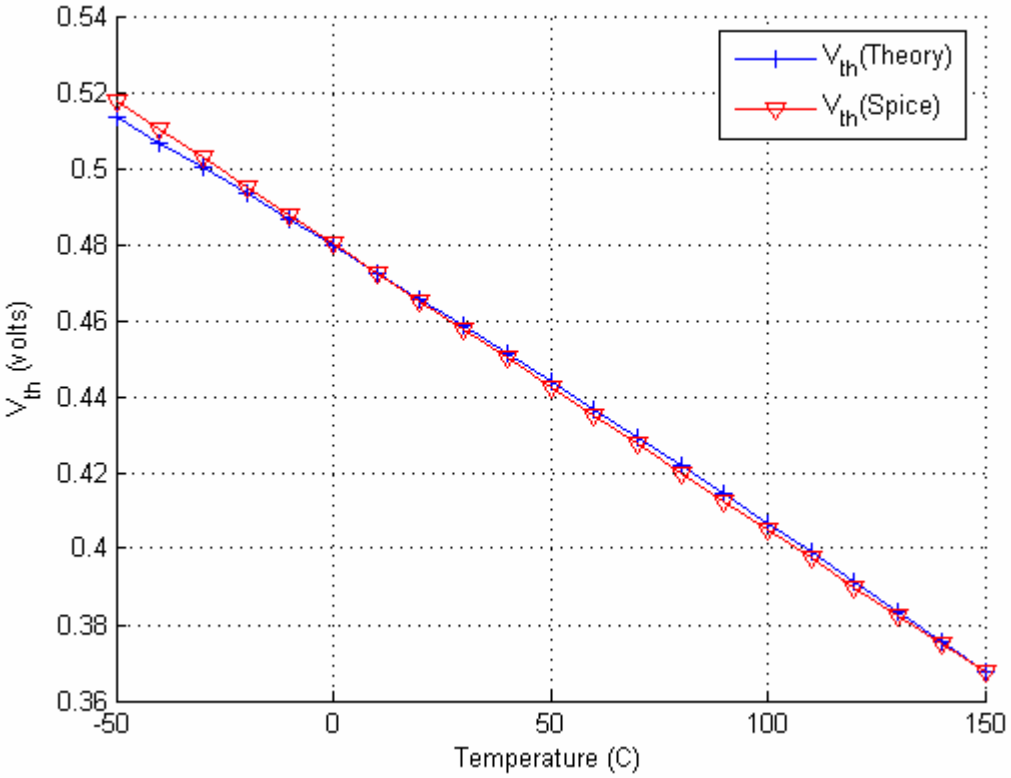

Figure 3.3: Threshold Voltage as a Function of Temperature

The temperature dependence is proportional to the temperature and the reasons are clear with further evaluation. By taking the derivative of V_{th} with respect to temperature

$$\frac{\partial V_{th}}{\partial T} = A + \frac{B}{\left[\frac{kT}{q}\left(2\ln\left(\frac{N_A}{K_{ni}}\right) - 3\ln(T)\right) + E_{go} - \frac{\alpha T^2}{T+\beta}\right]^{\frac{1}{2}}}, \tag{3.12}$$

where

$$A = \frac{k}{q}\ln\left(\frac{N_A}{N_{Dgate}}\right), \tag{3.13}$$

and

$$B = \frac{k}{q}\left(2\ln\left(\frac{N_A}{K_{ni}}\right) - 3\ln(T) - 3\right) - \frac{2\alpha T}{T+\beta} + \frac{\alpha T^2}{(T+\beta)^2}. \tag{3.14}$$

There is neither a dominant term nor a term that may be ignored. The least amount of error introduced by ignoring a term would contribute up to a 50% error over temperature. Therefore, this relationship may not be reduced further. The value of A is always negative when the condition $N_{gate} > N_A$ is true. Also, the value of B will remain negative as long as $K_{ni} > N_A$. In the case where both of these relationships are true, the threshold voltage is guaranteed to behave as a PTAT voltage.

Both N_A and N_{gate} are process parameters that change with scaling, but K_{ni} is a process independent parameter. It was pointed out in Chapter 2 that N_A will increase as the device technology scales downward. This will draw the relationship of $K_{ni} > N_A$ closer to failure. It was also mentioned in Chapter 2 that the gate technology may change in the future to allow for DTMOS devices. This will impact the ϕ_{ms} and ultimately the temperature characteristics of V_{th}. Once these types of devices are introduced, the ϕ_{ms} must be readjusted above. Presently, DTMOS devices are in the early stages of research and development, so the above analysis will hold true for years to come.

3.5 CARRIER MOBILITY

The effective mobility of a carrier, μ_{eff}, directly impacts the amount of drain current that a device may deliver, as shown in equation 2.1. It can be shown that the mobility is dependent on the strength of the electric fields, the impurity density, and temperature [11, 40, 41, 52-55]. The mobility's dependence on the impurity density may be calculated using the following formula that was developed by matching the measured data [55]:

$$\mu_{extrinsic} = \mu_{\min} + \frac{\mu_{\max} - \mu_{\min}}{1 + \left(\frac{N}{N_{ref}}\right)^{\alpha}}. \tag{3.15}$$

The parameters to be used for holes and electrons are shown in Table 3.2.

Table 3.2: Parameters for Determining Non-Intrinsic Silicon Mobility

	Electrons	Holes	Units
$\mu_{\min}$	92	47.7	$\frac{cm^2}{V \cdot s}$
$\mu_{\max}$	1360	495	$\frac{cm^2}{V \cdot s}$
N_{ref}	$1.3 \cdot 10^{17}$	$6.3 \cdot 10^{16}$	cm^{-3}
α	0.91	0.76	

The temperature dependence of the mobility is fairly large and may be generalized as

$$\mu(T) = \mu T^{-\gamma}, \tag{3.16}$$

where the temperature coefficient, γ, is theoretically 1.5 for an intrinsic material. However, γ has been shown to be as high as 2.42 to match experimental data [11, 55]. Combining equations 3.9, 3.10, and 2.59, the following mobility relationship is obtained:

$$\mu_{eff} = \frac{\mu_{extrinsic}\left(\frac{T}{T_{room}}\right)^{-\gamma}}{1 + \theta\left(V_{gs} - V_{th}\right)}; \tag{3.17}$$

The temperature was normalized to room temperature because the initial value of $\mu_{extrinsic}$ is calculated at room temperature. A plot of μ_{eff} from equation 3.11 and Smartspice is shown in Figure 3.4. The value of θ is calculated using equation 2.60 and the data for V_{gs} and V_{th} used in equation 3.11 is from the Smartspice simulation. As can be observed in the figure, there is a very good match between the theory and Spice simulation. With a

slight adjustment of the temperature dependent coefficient γ, an exact match is almost made between the theory and Spice.

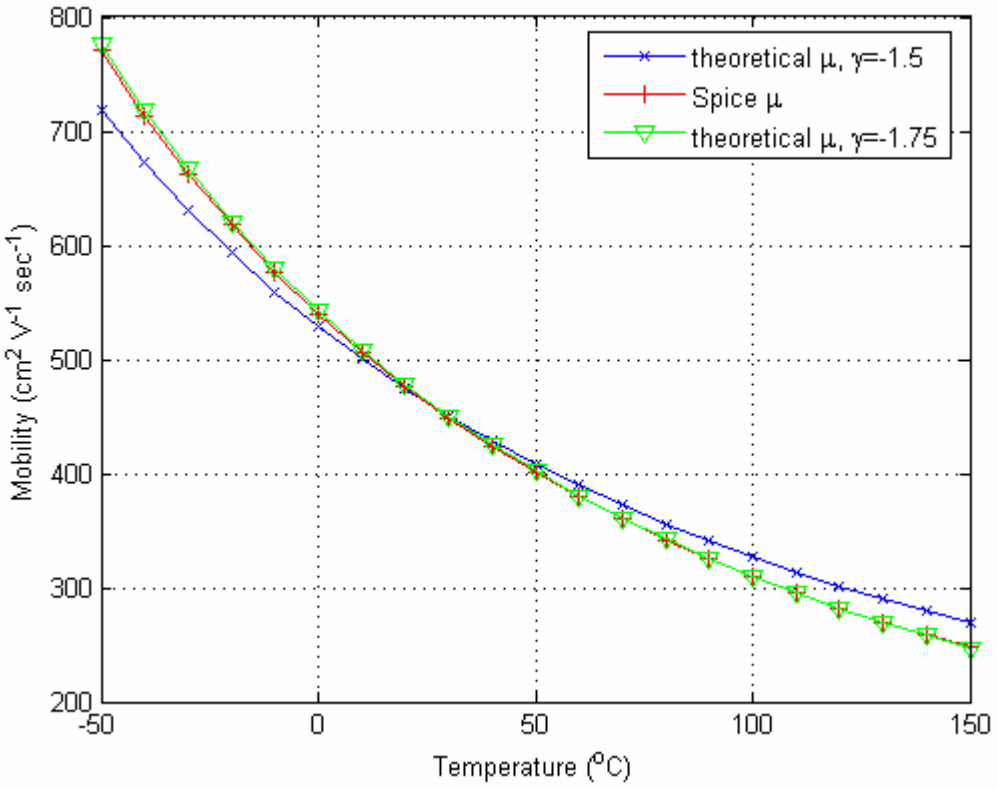

Figure 3.4: Mobility as a Function of Temperature

3.6 CONCLUSION

The temperature dependent parameters that affect MOSFETs were investigated. In some cases the temperature dependent model was forced to use equations that were curve fit to match experimental data. In other cases, the theory was based on the physics of the device. All of the temperature dependent parameters are based on the basic physical parameters such as the doping concentration of the device. Consistent results for the high-end device modeling software and the theory were obtained. With the current-voltage characteristic equations of MOSFETs being known

and depending on these basic temperature dependent parameters, circuits made entirely of CMOS devices may be developed to generate PTAT and CTAT responses.

Chapter 4: CTAT and PTAT Generating Circuits

4.1 INTRODUCTION

In this chapter several circuits are presented that are based on the temperature characteristics of a MOSFET. PTAT and CTAT voltages or currents are generated that can potentially be summed together to produce a temperature stable voltage reference. In order to evaluate the temperature response of the device in a manner that brings manageable insight to the circuit characteristics, the equivalents of LEVEL1 to LEVEL3 models are used [11, 12, 15, 57-61]. These device modeling equations maintain an accuracy level that allows for an understanding of the temperature dependent characteristics of a given circuit. It is important to note that these equations are applicable to any CMOS process that exists in any manufacturing facility. Once a circuit is developed, it may then be simulated. The simulation tool uses more complex equations and will produce results that will be accurate for the manufactured circuit. All devices described in this chapter are assumed to have long channels in order to control the short channel effects. The channel lengths, for example, may be on the order of $2\mu m$ in a $90nm$ process. The temperature range that is used for discussion in this chapter is from $-50^{o}C$ to $150^{o}C$, unless specified otherwise.

4.2 CTAT SOURCES

With a constant current being driven into a diode-connected MOSFET device, as shown in Figure 4.1, the V_{gs} voltage will have a CTAT response. First, the temperature dependence in the saturation region will be discussed. Then, the temperature dependence in the subthreshold region will be discussed. In the saturation region, equation 2.1 may

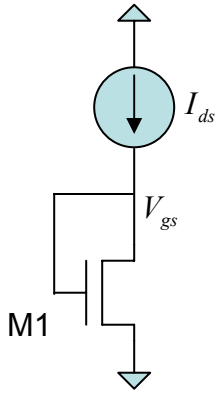

Figure 4.1: Diode-connected MOSFET

be used. Assuming that $f\left(V_{gs}, V_{th}, R_s, v_{sat}, N_A, V_{ds}, \lambda\right)$ is equal to one (i.e. for an ideal device) and solving for V_{gs},

$$V_{gs} = \sqrt{\frac{2I_d L_{eff}}{\mu_{eff} C_{ox} W_{eff}}} + V_{th}\,. \tag{4.1}$$

This equation may be expanded to show its temperature dependence. By substituting in the expressions for mobility and threshold voltage from Chapter 3,

$$V_{gs} = \sqrt{\frac{2I_d}{\mu_{eff}(T_{nom}) C_{ox} S}}\left(\frac{T}{T_{nom}}\right)^{\frac{\gamma}{2}} + V_{th\_tnom} + K_{vth}\left(T - T_{nom}\right), \tag{4.2}$$

where

$$S = \frac{W_{eff}}{L_{eff}}\,. \tag{4.3}$$

Taking the derivative of V_{gs} with respect to the temperature results in

$$\frac{\partial V_{gs}}{\partial T} = \frac{\gamma}{2}\sqrt{\frac{2I_d}{\mu_{eff}(T_{nom}) C_{ox} S}}\left(\frac{T}{T_{nom}}\right)^{\frac{-\gamma}{2}} + K_{vth}\,. \tag{4.4}$$

As mentioned in Chapter 3, K_{vth} is negative, so if

$$-K_{vth} > \frac{\gamma}{2}\left(\frac{1}{T_{nom}}\right)^{\frac{\gamma}{2}}\sqrt{\frac{2I_d}{\mu_{eff}(T_{nom}) C_{ox} S}}\left(\frac{T_{\min}}{T_{nom}}\right)^{\frac{-\gamma}{2}}, \tag{4.5}$$

then the temperature characteristics of a diode-connected MOSFET will exhibit a negative temperature coefficient. The newly introduced parameter, $T_{\min}$, is the minimum temperature in the temperature range of interest.

With a large I_d current and small S value it is possible to violate the expression in equation 4.5. This scenario is not realistic in typical applications and there are better ways to generate a PTAT voltage. Typical circuit implementations need well matched devices and low power. Therefore, the parameter S is typically large while the drain current I_d is small. As an example, by choosing $S = 50$ and $I_d = 100uA$ while using the temperature dependent parameters from Chapter 3, it can be shown that the relationship requirements of equation 4.5 are satisfied by a couple orders of magnitude (i.e. $-\left(-0.78\frac{mV}{^oC}\right) > 5.07\frac{uV}{^oC}$). Figure 4.2 shows a plot of V_{gs} and the derivative of V_{gs} with respect to temperature for a device operating in the saturation region. These plots were generated using equations 4.2 and 4.4 where the parameter values of $S = 2$ and $I_d = 25\mu A$ were chosen.

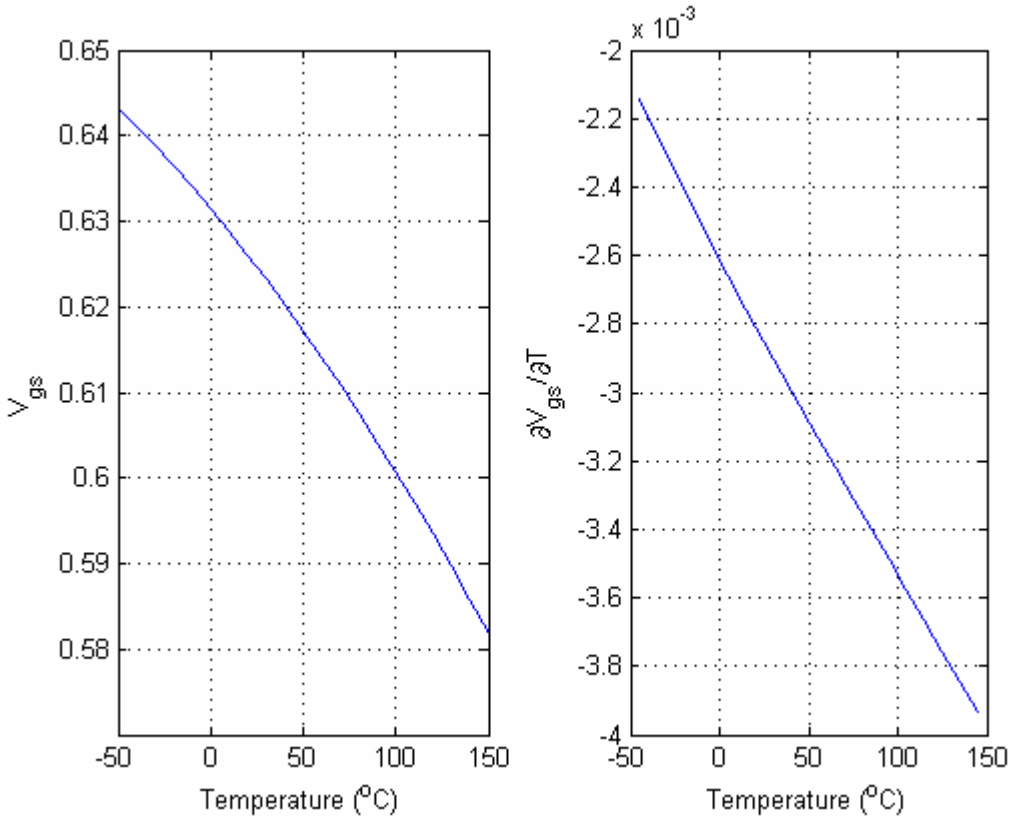

Figure 4.2: V_{gs} and $\frac{\partial V_{gs}}{\partial T}$ as a Function of Temperature in the Saturation Region

In the same configuration shown in Figure 4.1, the device may operate in the subthreshold region. In this case, the gate voltage is reduced to a low enough value that the surface charge is actually operating in the weak inversion region. A simple and accurate relationship that has been developed is

$$V_{gs} = n\phi_t \ln\left(\frac{I_d}{I_s}\right) + V_{th} \text{ [12, 56-61]}, \tag{4.6}$$

where I_s is the specific current and n is the subthreshold slope. All of the parameters, except for I_d, are temperature dependent. The subthreshold slope is temperature dependent due to the depletion capacitance, where

$$n = 1 + \frac{C_{dep}}{C_{ox}}. \tag{4.7}$$

This expression may be expanded to show its temperature dependent parameters:

$$n = 1 + \frac{\sqrt{q\varepsilon_{si}N_A}}{C_{ox}\sqrt{2\left(2\phi_t \ln\left(\frac{N_A}{n_i}\right) - 3\phi_t\right)}}. \tag{4.8}$$

The temperature dependence of n is reasonably small over the temperature span. It will vary from approximately 1.21 to 1.28. This variation will not contribute more than a 10% error. For modeling simplification, it will be assumed to be a constant value. The specific current is defined as

$$I_s = 2n\mu_{eff}C_{ox}S\phi_t^{\,2}. \tag{4.9}$$

By expanding all of the temperature dependent parameters and solving for V_{gs} in equation 4.6, it is found that

$$V_{gs} = n\frac{kT}{q}\left[\ln(C_4) + (\gamma - 2)\ln(T)\right] + V_{th\_tnom} + K_{vth}(T - T_{nom}), \tag{4.10}$$

where the temperature independent parameter

$$C_4 = \frac{I_d q^2}{2n\mu(T_{nom})C_{ox}Sk^2T_{nom}^{\gamma}}. \tag{4.11}$$

Setting the parameters $S = 50$ and $I_d = 5\mu A$, a plot of V_{gs} with respect to temperature is shown in Figure 4.3.

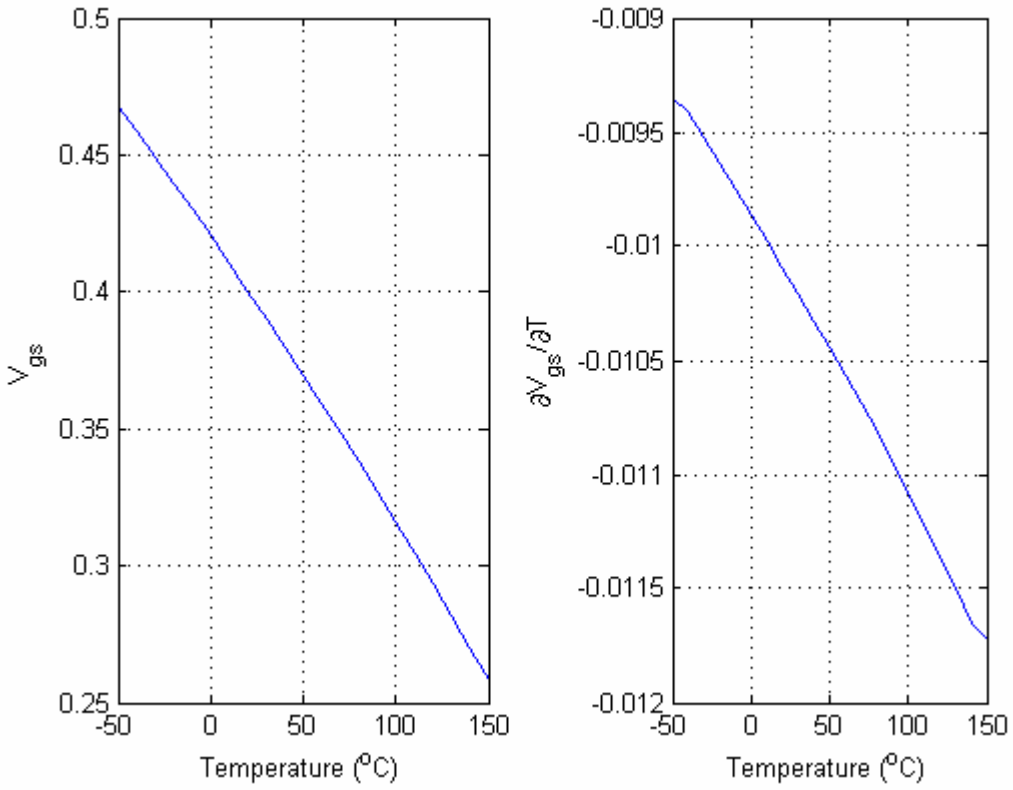

Figure 4.3: V_{gs} and $\frac{\partial V_{gs}}{\partial T}$ as a Function of Temperature in the Subthreshold Region

As illustrated in Figure 4.3, V_{gs} is dominated by a negative temperature coefficient. Taking the derivative of V_{gs} with respect to temperature results in the following expression:

$$\frac{\partial V_{gs}}{\partial T} = n\frac{k}{q}\left[\ln(C_4) + (\gamma - 2)\ln(T)\right] + n\frac{k}{q}(\gamma - 2) + K_{vth}. \tag{4.12}$$

As long as the relationship of

$$-K_{vth} > n\frac{k}{q}\left[\ln(C_4) + (\gamma - 2)\ln(T)\right] + n\frac{k}{q}(\gamma - 2) \tag{4.13}$$

is satisfied, then V_{gs} will have a negative temperature coefficient. Operating in the subthreshold region requires that the current to device size ratio is very small. This

combined with the temperature coefficient dependencies of V_{th}, makes it nearly impossible for V_{gs} to not have a negative temperature coefficient.

Clearly, the threshold voltage dominates the CTAT characteristics of the MOSFET device. Figure 4.4 shows a circuit method that may be used to extract the

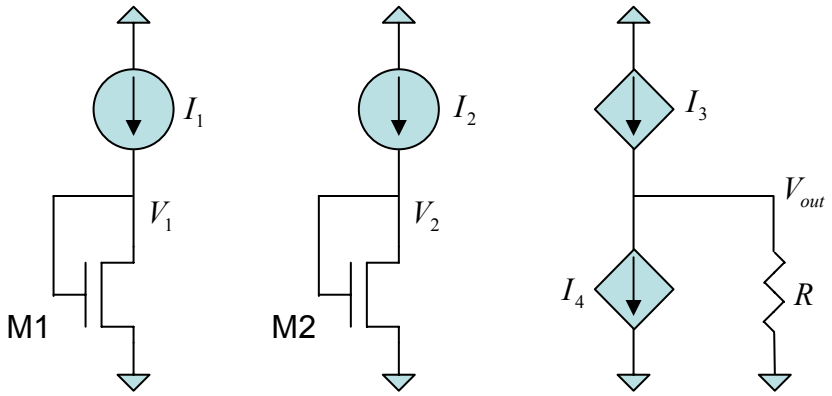

Figure 4.4: Circuit Diagram to Extract V_{th}

threshold voltage [62]. Assuming that both $M1$ and $M2$ are in the saturation region and that

$$\eta = \sqrt{\frac{I_1 S_2}{I_2 S_1}}, \tag{4.15}$$

the threshold voltage may be extracted by implementing the following equation:

$$V_{th} = \frac{V_1 - \eta V_2}{1 - \eta}. \tag{4.16}$$

One solution is to force $\eta = 2$ by allowing $I_1 = 4I_2$ and $S_1 = S_2$ so that the threshold voltage may be extracted by

$$V_{th} = 2V_2 - V_1. \tag{4.17}$$

The voltage controlled current sources I_3 and I_4 are easily implemented and are used in the voltage reference circuit presented in Chapter 6. By implementing

$$I_3 = \frac{2V_2}{R} \text{ and} \tag{4.18}$$

$$I_4 = \frac{V_1}{R}, \tag{4.19}$$

where R is a resistor used in the voltage to current conversion circuit, the resulting output voltage V_{out} will have implemented equation 4.17 and is equal to the threshold voltage of the devices.

Three methods have been presented that may be used to extract a CTAT voltage. There are other methods that may prove to be successful, such as developing a circuit that extracts the temperature dependence of mobility into a voltage. These circuits were developed by using simplified I_d-V_{gs} relationships.

4.3 PTAT SOURCES

Various circuits can generate a PTAT voltage or PTAT current source. Several feasible architectures will be presented. First, a few circuit architectures are shown that rely on the voltage differences of two diode-connected devices that have different current densities. Second, a circuit that has one device in saturation and a second device operating in the linear region is shown. Third, a circuit that current mirrors an input voltage into a diode connected device is shown. All of these circuits result in a PTAT voltage or PTAT current reference.

The difference of the gate to source voltage, ΔV_{gs}, of two devices having different current densities and both devices operating in the saturation region may be derived by following these expressions:

$$\Delta V_{gs} = V_{gs1} - V_{gs2}; \tag{4.20}$$

$$V_{gs1} = \sqrt{\frac{2I_{d1}}{\mu(T_{nom})C_{ox}S_1}\left(\frac{T}{T_{nom}}\right)^{\gamma}} + V_{th}; \tag{4.21}$$

$$V_{gs2} = \sqrt{\frac{2I_{d2}}{\mu(T_{nom})C_{ox}S_2}\left(\frac{T}{T_{nom}}\right)^{\gamma}} + V_{th}; \tag{4.22}$$

$$\Delta V_{gs} = \sqrt{\frac{2}{\mu(T_{nom})C_{ox}T_{nom}^{\gamma}}}\left[\sqrt{\frac{I_{d1}}{S_1}} - \sqrt{\frac{I_{d2}}{S_2}}\right]T^{\frac{\gamma}{2}}. \tag{4.23}$$

Assuming that the current density of device $M1$ is greater then the current density of device $M2$, then ΔV_{gs} will respond with a temperature dependence related to the mobility temperature dependent coefficient.

If the devices are operating in the subthreshold region, then the results are significantly modified. A PTAT voltage based on ΔV_{gs} of devices operating in the subthreshold region may be obtained by following these equations and using proper substitutions:

$$\Delta V_{gs} = V_{gs1} - V_{gs2}; \tag{4.24}$$

$$V_{gs1} = n\phi_t \ln\left(\frac{I_{d1}}{I_{s1}}\right) + V_{th}; \tag{4.25}$$

$$V_{gs2} = n\phi_t \ln\left(\frac{I_{d2}}{I_{s2}}\right) + V_{th}; \tag{4.26}$$

$$\Delta V_{gs} = n\frac{kT}{q}\ln\left(\frac{I_{d1}S_2}{I_{d2}S_1}\right). \tag{4.27}$$

In this case the voltage is directly proportional to temperature while all other device temperature characteristics cancel each other out. This is a favorable property because the second order temperature effects should be minimal and related to device mismatching parameters and the subthreshold slope n. It is assumed that the temperature coefficient of a resistor is negligible or will be cancelled out in the implementation of a temperature stable reference. Figures 4.5 and 4.6 show two circuits that generate a PTAT current reference based on the ΔV_{gs} of two devices.

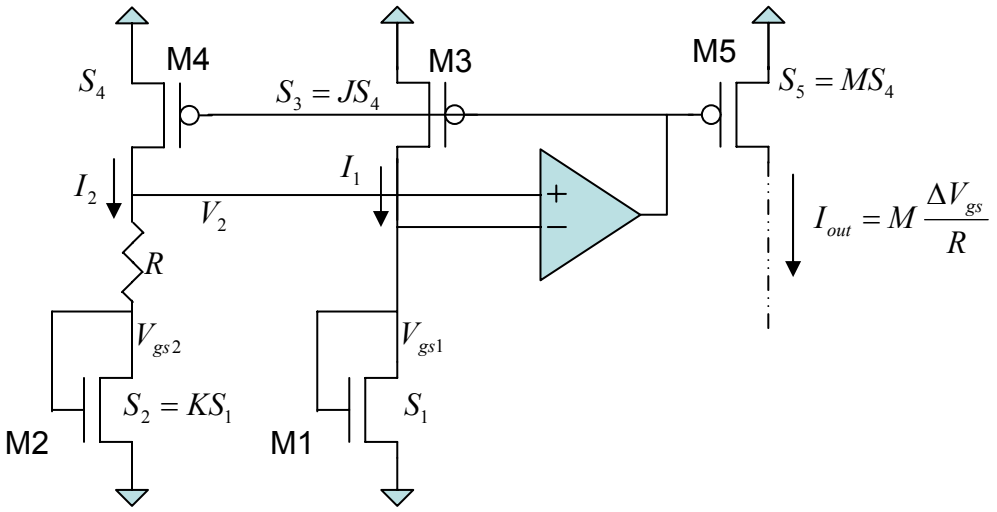

Figure 4.5: PTAT Current Generator with OPAMP

In both of the circuits shown in Figures 4.5 and 4.6, there exists a potential startup problem with the circuit. Therefore additional circuits must be added to ensure the proper operation. The temperature slope of the PTAT current is adjusted by the parameters M, J, and K, which are all easy layout adjustments. This analysis is completed in the follow section.

With the assumption that $M1$ and $M2$ are operating in the subthreshold region in Figure 4.5, the PTAT output current I_{out} may be found [63]. The OPAMP forces the node voltages

$$V_2 = V_{gs1}. \tag{4.28}$$

Using KVL, it is determined that

$$V_2 = V_{gs2} + I_2 R. \tag{4.29}$$

By solving equation 4.29 for I_2 and substituting in equation 4.24

$$I_2 = \frac{\Delta V_{gs}}{R}. \tag{4.30}$$

Based on the p-channel device size differences, I_{out} and I_1 are related to I_2 in the following manner:

$$I_{out} = MI_2; \tag{4.31}$$

$$I_1 = JI_2. \tag{4.32}$$

The size differences between $M1$ and $M2$ are related by

$$S_2 = KS_1. \tag{4.33}$$

By substituting equations 4.33 and 4.32 into equation 4.27, it is found that

$$\Delta V_{gs} = n\frac{kT}{q}\ln(JK). \tag{4.34}$$

Now, the output current may be solved by substituting equations 4.31 and 4.34 into equation 4.30 resulting in the expression

$$I_{out} = Mn\frac{kT}{qR}\ln(JK). \tag{4.35}$$

The circuit in Figure 4.5 clearly produces a PTAT current source. However, because today's applications strive to minimize area and power consumption, the OPAMP is not desirable.

A similar result for a reference current may be obtained with the circuit presented in Figure 4.6 [15]. In this case, the operational amplifier is not required.

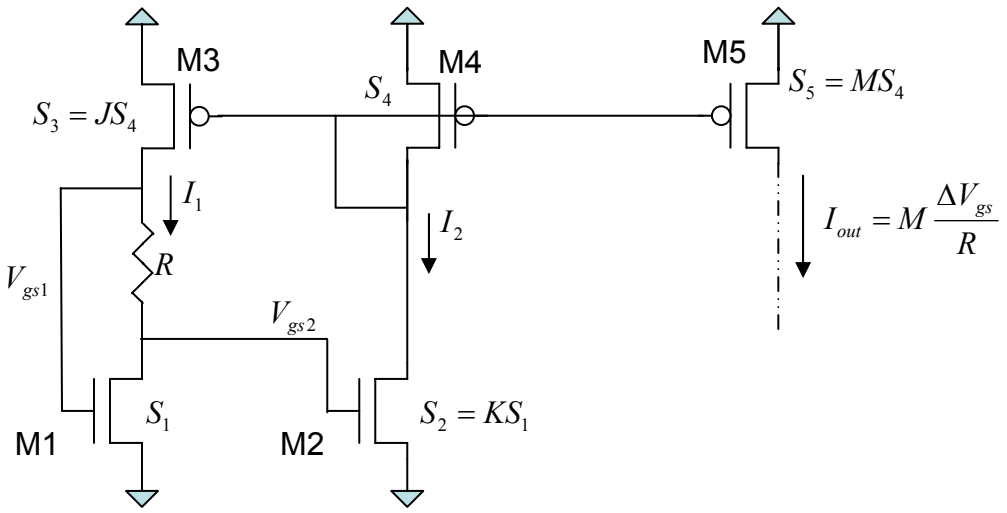

Figure 4.6: PTAT Current Generator without OPAMP

Completing the KVL to relate both of the gate to source voltages of $M1$ and $M2$ to each other, results in the expression

$$V_{gs1} = V_{gs2} + I_1 R . \tag{4.36}$$

The relationships of I_1 to I_2 and of I_{out} to I_1 may be expressed as follows:

$$I_1 = JI_2 ; \tag{4.37}$$

$$I_{out} = MI_1 . \tag{4.38}$$

Solving for I_1 in equation 4.36 and substituting the result into equation 4.38 results in

$$I_{out} = M \frac{\Delta V_{gs}}{R} . \tag{4.39}$$

By expanding ΔV_{gs} and using the area ratio between devices $M1$ and $M2$, the final result is

$$I_{out} = Mn \frac{kT}{qR} \ln(JK). \tag{4.40}$$

This result is identical to the results of the circuit presented in Figure 4.5.

The circuit shown in Figure 4.7 also results in a PTAT voltage output [65, 66]. The analysis is more complex than with the previous PTAT solutions. Assuming that the input voltage is a constant, the device $M2$ will operate in the saturation region while device $M1$ operates in the linear region. By allowing the

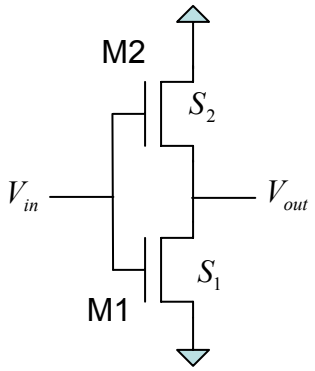

Figure 4.7: PTAT Voltage Generator

parameter β to be defined as

$$\beta = \mu C_{ox} \frac{S}{2}, \tag{4.31}$$

the saturation current relationship for device $M2$ may then be written as

$$I_{d2} = \beta_2 \left(V_{gs2} - V_{th}\right)^2. \tag{4.32}$$

The linear current relationship for device $M1$ may be expressed as

$$I_{d1} = 2\beta_1 \left[\left(V_{gs1} - V_{th}\right) V_{ds1} - \frac{1}{2} V_{ds1}^2 \right]. \tag{4.33}$$

By inspection of the circuit in Figure 4.7, it is found that the following expressions are valid:

$$V_{gs1} = V_{in}; \tag{4.34}$$

$$V_{gs2} = V_{in} - V_{ds1} \tag{4.35}$$

$$V_{ds1} = V_{out}. \tag{4.36}$$

Substituting of all of these parameters and knowing that the currents must be equal, the following expression may be written:

$$2\beta_1 \left[\left(V_{in} - V_{th}\right) V_{out} - \frac{1}{2} V_{out}^2 \right] = \beta_2 \left(V_{in} - V_{out} - V_{th}\right)^2. \tag{4.37}$$

The following parameters may be defined as

$$C = \frac{\beta_2}{2\beta_1} \tag{4.38}$$

and

$$K = \left(V_{in} - V_{th}\right), \tag{4.39}$$

equation 4.37 may be expanded and rewritten into the quadratic equation

$$\left(\frac{1}{2} + C\right) V_{out}^2 + \left(-K - 2CK\right) V_{out} + CK^2 = 0. \tag{4.40}$$

Solving this equation results in the following general solution

$$V_{out} = \left(V_{in} - V_{th}\right)\left[1 \pm \frac{1}{\sqrt{\frac{\beta_2}{\beta_1} + 1}}\right]. \tag{4.41}$$

Upon further inspection and by the use of Spice, the root that uses the positive sign is not a valid solution so equation 4.41 becomes

$$V_{out} = \left(V_{in} - V_{th}\right)\left[1 - \frac{1}{\sqrt{\frac{\beta_2}{\beta_1} + 1}}\right]. \tag{4.42}$$

In evaluating which root was valid, an error of less than 0.1% was obtained when using the Spice Level1 models. The temperature dependence due to mobility is cancelled and the temperature dependence of V_{out} is related to a constant minus the threshold voltage. As the threshold voltage decreases over temperature, the output voltage will increase.

There is one concern about the output voltage of this circuit. Due to the nature of the power supplies and how they are being reduced with the technology scaling, the output voltage may be very small. It is potentially a difficult problem to level shift this voltage or convert it into a current. Another concern with the presented circuit may be the error introduced due to the bulk potentials being different for both devices. This can be alleviated by converting the n-channel devices into p-channel devices and connecting the bulk ties to the drain of the devices.

The circuit shown in Figure 4.8 may be setup to output a CTAT or PTAT voltage. Both cases of operation in the saturation region and subthreshold region for the devices $M1$ and $M2$ are investigated. Generally speaking, for both cases, the

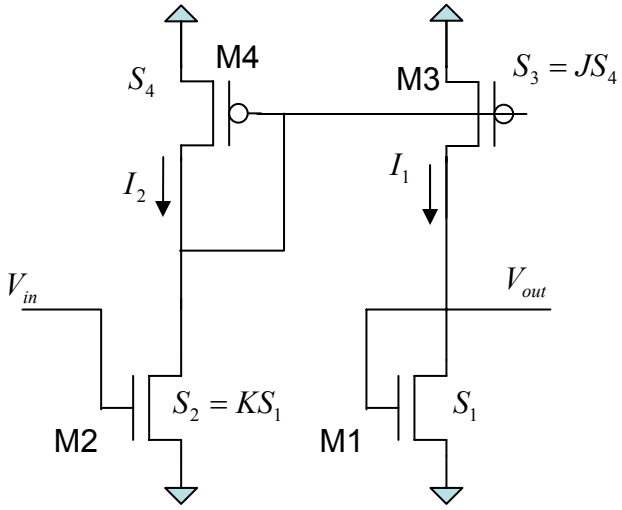

Figure 4.8: Current Mirror into a Diode-connected Device

currents of $M1$ and $M2$ are related by

$$I_1 = JI_2 . \tag{4.42}$$

Let the parameter C be defined as

$$C = \sqrt{JK} . \tag{4.43}$$

The current mirror consisting of M3 and M4 requires that

$$\beta_1 \left(V_{out} - V_{th}\right)^2 = J\beta_2 \left(V_{in} - V_{th}\right)^2 . \tag{4.44}$$

Solving for V_{out}, the expression for the devices in saturation is

$$V_{out} = CV_{in} + \left(1 - C\right)V_{th} . \tag{4.45}$$

Equations 4.44 and 4.45 may also be written for the case where the devices M1 and M2 are operating in the subthreshold region. The result is

$$I_{s1} \exp\left(\frac{V_{out} - V_{th}}{n\phi_t}\right) = JI_{s2} \exp\left(\frac{V_{in} - V_{th}}{n\phi_t}\right). \tag{4.46}$$

Solving this expression for V_{out} and substituting in the parameter C, gives the result

$$V_{out} = V_{in} - 2n\phi_t \ln\left(C\right). \tag{4.47}$$

By inspection of equations 4.45 and 4.47, if $C > 1$ it is clear that the output voltage will act as a PTAT voltage. Otherwise, the circuit will generate a CTAT voltage.

4.4 PIECEWISE LINEAR CORRECTION CIRCUIT

The output of a properly summed CTAT and PTAT current will result in a reference voltage that is reasonably stable with respect to temperature. There remains, however, some residual error. Piecewise Linear (PWL) circuits may be used to further reduce this error. From the CTAT-PTAT sum, a PWL circuit is a circuit for which the output current is active only above or below a specified temperature. The specifics of the process used to reduce the residual error will become clearer in Chapter 5.

A PWL circuit that will generate a current above a specified temperature is shown in Figure 4.9 [64]. It includes two current sources reflecting scaled versions of

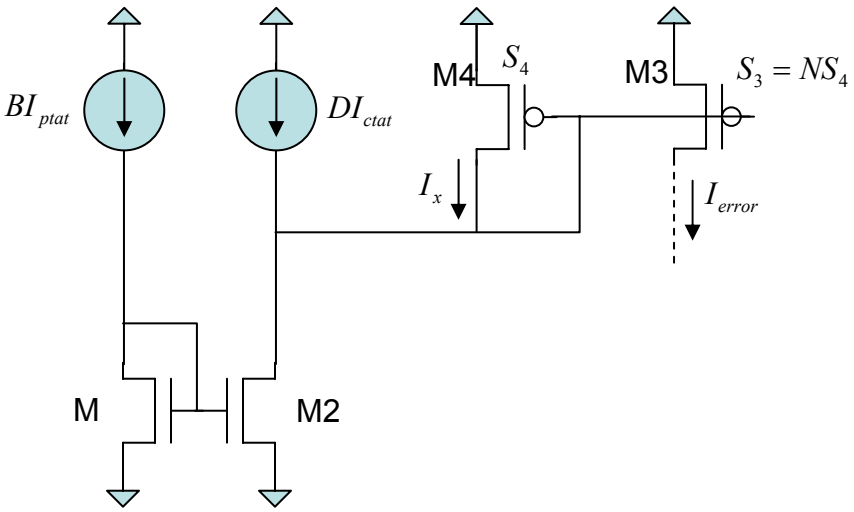

Figure 4.9: PWL Circuit Active when $T > T_{break}$

CTAT and PTAT currents, a method to take the difference between these two currents, and a method to drive a scaled current based on the difference of the two currents. Even though the PWL name would indicate that it has a linear response, the response is

actually highly dependent on the CTAT and PTAT signals. However, due to the naming convention of Rincon-Mora [64], the PWL name is used throughout this document.

The parameters B and D allow the I_{ptat} and I_{ctat} to be scaled, respectfully. The devices $M1$ and $M2$ mirror the scaled I_{ptat} current so that the difference of the temperature dependent currents may be taken. With device $M4$ being diode-connected and tied into the same node which takes the difference of the temperature dependent currents, the current through $M4$ may be expressed as

$$I_x = BI_{ptat} - DI_{ctat}. \tag{4.48}$$

Based on the current mirror to produce the scaled current difference of these currents, it is found that

$$I_{error} = N\left(BI_{ptat} - DI_{ctat}\right). \tag{4.49}$$

With further evaluation, it should be apparent that I_{out} will not produce any current until the condition

$$BI_{ptat} - DI_{ctat} > 0 \tag{4.50}$$

The breaking point at which I_{error} will start to conduct is the temperature at which the scaled temperature dependent currents are equal. This is under the condition $BI_{ptat}\left(T_{break}\right) = DI_{ctat}\left(T_{break}\right)$, where T_{break} is the desired temperature to begin conducting current. The scaling relationship between D and B will typically not be an integer value and the parameters only relate to each other. Therefore, the parameter D may be chosen by the designer, which leaves the following relationship to be solved:

$$B = D\frac{I_{ctat}\left(T_{break}\right)}{I_{ptat}\left(T_{break}\right)}. \tag{4.51}$$

An example of this PWL circuit that conducts current above a specified temperature is shown below. The following temperature dependent equations were used:

$$I_{ctat} = 90\cdot 10^{-6} - 170\cdot 10^{-9}\cdot\left(T - T_{nom}\right); \tag{4.52}$$

$$I_{ptat} = 52\cdot 10^{-6} + 56.6\cdot 10^{-9}\cdot\left(T - T_{nom}\right). \tag{4.53}$$

The temperature T_{nom} is room temperature in degrees Kelvin. By choosing the breaking point to occur at $50^{o}C$ and choosing the scaling constant D to be equal to one, the scaling constant B is found to be equal to 1.73. Figure 4.10 plots the temperature response of this example using the assumptions above and equations 4.49, 4.52, and 4.53.

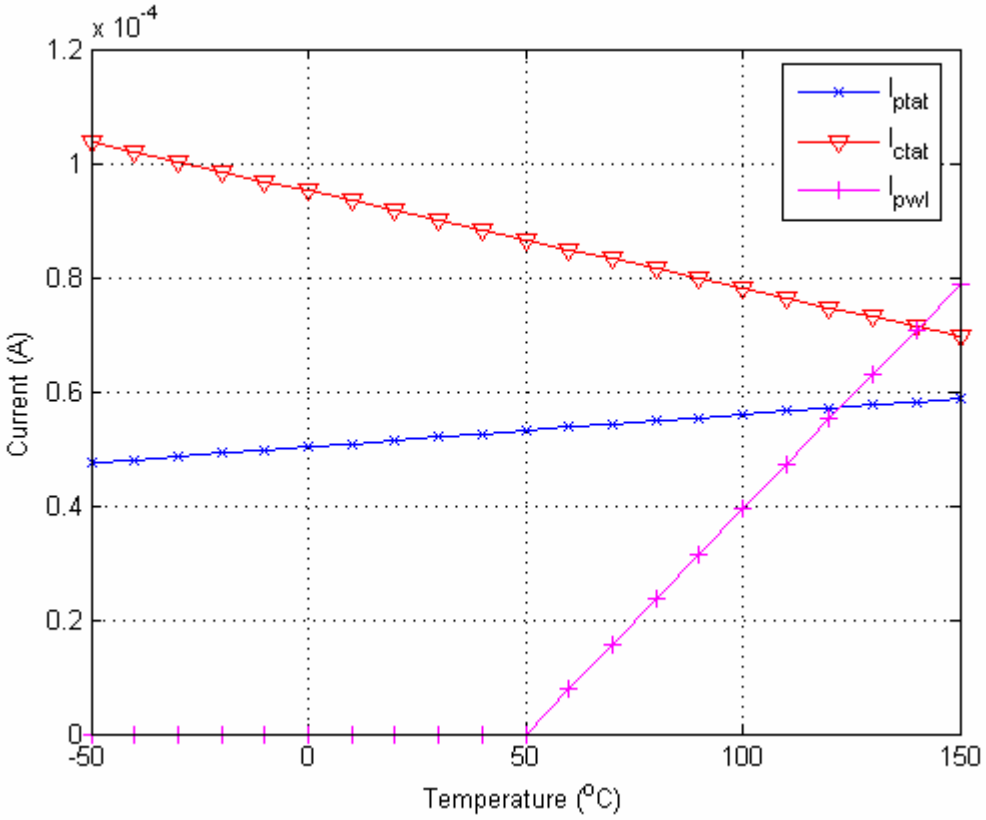

Figure 4.10: PWL Current that Responds with $T > T_{break}$

In order to change the breaking point, the ratio of $B:D$ has to change. The slope of the PWL output current is then set by the parameter N, which was set to three in the above plot.

A second PWL circuit may be designed that conducts current below a specified temperature. Such a circuit is shown in figure 4.11. This circuit is primarily the same as the one in Figure 4.9, except that the

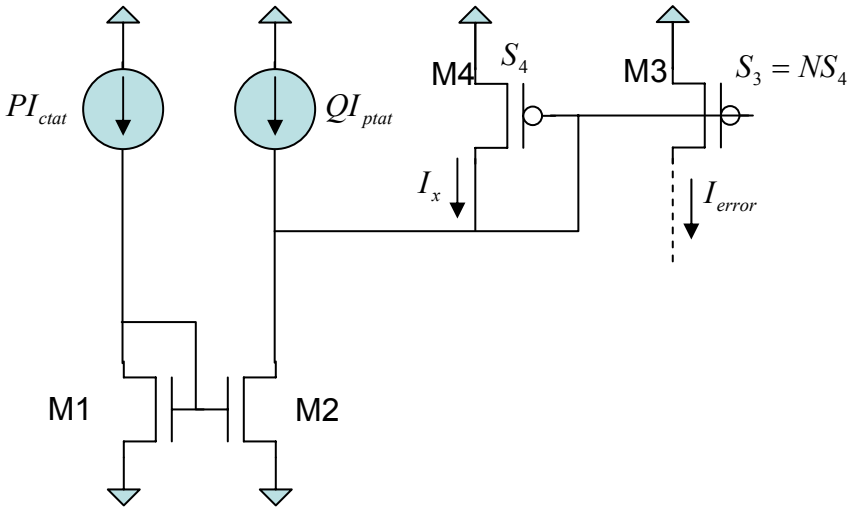

Figure 4.11: PWL Circuit Active when $T < T_1$

current sources for the PTAT and CTAT have been swapped. This circuit will not produce an output current above the specified output temperature T_{break}. In this case, using the same procedure as above, the output current

$$I_{error} = N\left(PI_{ctat} - QI_{ptat}\right). \tag{4.54}$$

The PWL output current will be conducting until the relationship

$$PI_{ctat} - QI_{ptat} > 0 \tag{4.55}$$

is violated. At that particular temperature and above, no I_{error} current will be produced. To solve for the scaling coefficients to set the temperature breakpoint, the relationship of

$$P = Q\frac{I_{ptat}\left(T_{break}\right)}{I_{ctat}\left(T_{break}\right)} \tag{4.55}$$

may be used. Using the same temperature dependent equations for I_{ctat} and I_{ptat} as above, and assuming Q is equal to one, a solution to equation 4.55 may be obtained. In one variation, the temperature T_{break} is set to $70^o C$. This results in the scaling constant P to be equal to 1.53. With the parameter N equal 3, the results are shown in Figure

4.12. The temperature at which I_{error} goes to zero may be adjusted by changing the $P:Q$ ratio, while the slope of PWL output current can be adjusted by adjusting N.

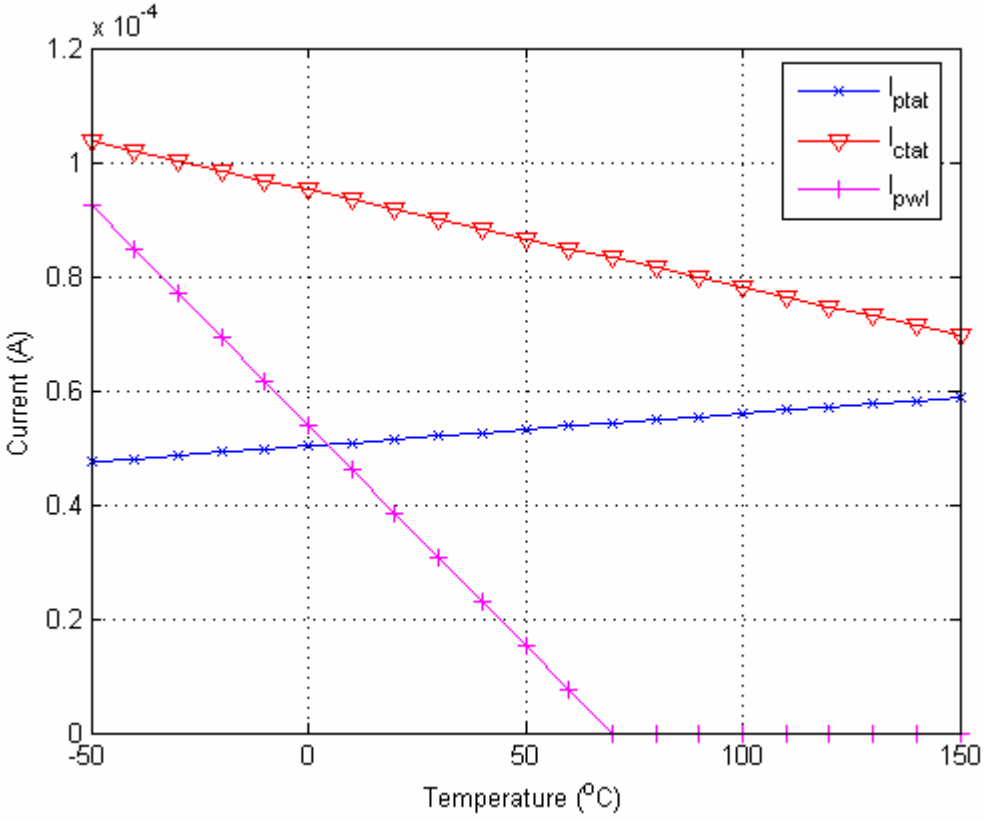

Figure 4.12: PWL Current that Responds with $T < T_{break}$

4.5 CONCLUSION

Several circuits were proposed that have the capability to generate a voltage or current with a PTAT or CTAT response. The PWL circuits provide a method in which a PTAT or CTAT current may be generated that operates above or below a specified temperature. These circuits may be combined in such a way to produce a voltage or current reference that is quite stable with respect to temperature. There are a significant number of circuits that may be used to produce these stable sources. Ideally, the minimum number of stacked devices is used. This will allow for the voltage reference

technology to operate at the minimum power supply, which will enable the voltage reference architecture to survive as the process technology is scaled.

Chapter 5: Voltage Reference System Level

5.1 INTRODUCTION

In this chapter a voltage reference architecture is shown that generates a stable reference by summing multiple temperature dependent currents at a single node and converting it into a voltage. By representing each current generated by these blocks as a polynomial and applying the Method of Least Squares while doing an iterative search for the best PWL temperature breakpoints, an optimized, higher order voltage reference may be designed at the system level. Chapter 6 uses the methodology presented in this chapter to combine several circuits that were presented in Chapter 4 to demonstrate the power of this methodology.

5.2 VOLTAGE REFERENCE ARCHITECTURE

The voltage reference architecture is made up of a PTAT, a CTAT, and multiple PWL circuits. Figure 5.1 shows a system level diagram of the architecture. As can be observed in Figure 5.1, a PTAT current, a CTAT current, and multiple PWL currents are scaled and summed together, resulting in a total current I_{total}. This total current is then forced into a resistor (i.e. scaled) to generate the temperature independent output voltage V_{ref}.

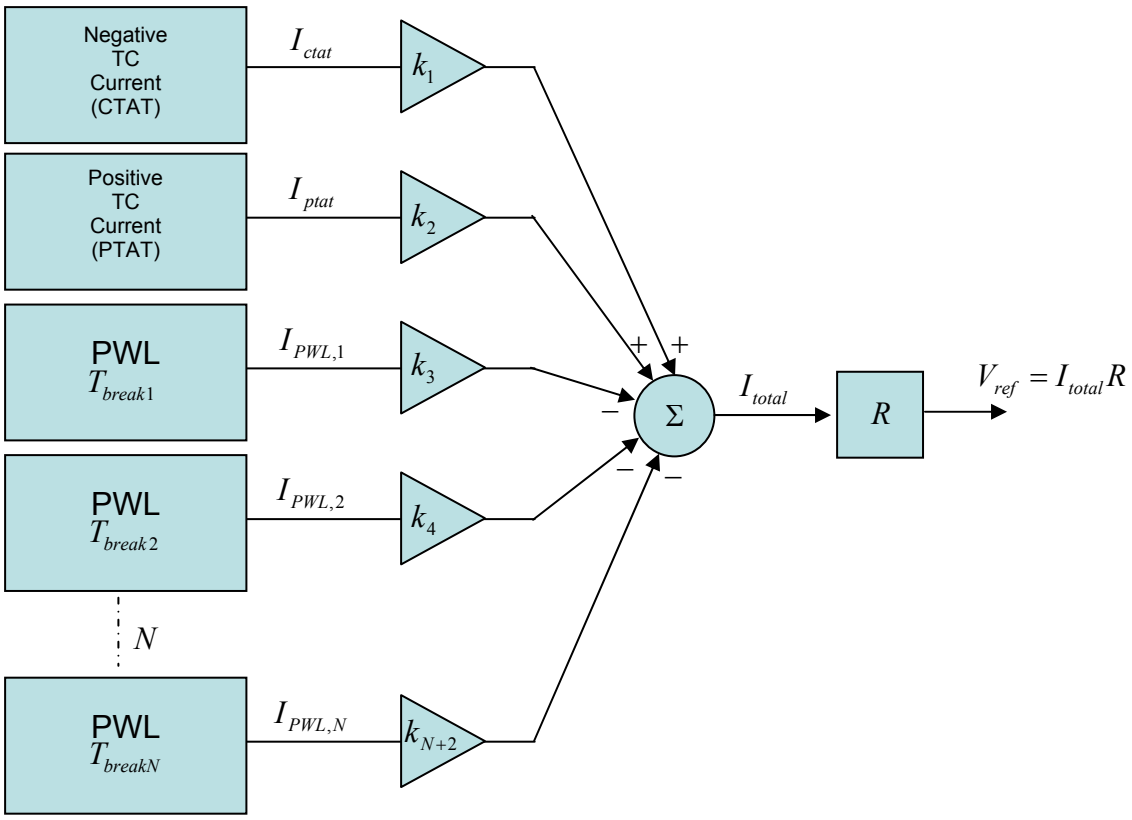

Figure 5.1: System Architecture of $(N+1)^{th}$ Order Voltage Reference

The core of the voltage reference is the scaled summation of the PTAT and CTAT currents, which may be referred to as I_{core}. The currents I_{ctat} and I_{ptat} are scaled by gains K_1 and K_2 and are summed into the summing node I_{total}. It was demonstrated in Chapter 4 that the temperature response of any silicon device will not be perfectly linear. Instead, the temperature response of each circuit is best described by a polynomial. Therefore the response of I_{core} should not be expected to be a constant over temperature. A number of PWL currents may be activated or deactivated at specific temperature breakpoints to force the total current to a minimum deviation over temperature. By summing up all of the polynomials, the output voltage of the reference has the general form of

$$V_{ref} = \left(K_1 I_{ctat} + K_2 I_{ptat} - K_3 I_{PWL,1} - K_4 I_{PWL,2} - \ldots - K_{N+2} I_{PWL,N}\right) R . \quad (5.1)$$

One of the goals of this architecture, besides initial accuracy, is to minimize the voltage variation over temperature. The Temperature Coefficient (TC) is a metric for voltage references and is expressed as [68, 69]:

$$TC = \frac{V_{\max} - V_{\min}}{V_{nom} T_{range}} . \quad (5.2)$$

The parameter V_{nom} is the mean of the reference voltage over the temperature range T_{range}. The voltages $V_{\max}$ and $V_{\min}$ are the maximum and minimum voltages that occur in the overall temperature range. This metric will be used consistently as a measure of performance.

5.3 METHOD OF LEAST SQUARES

The Method of Least Squares (MLS) is a way to best fit V_{ref} to a predetermined temperature response by solving for the scaling coefficients $K_1 - K_{N+2}$. In the case of a stable voltage reference, the best fit curve will be a constant voltage over temperature. In some applications, a linear curve with temperature may be preferred. By applying the Method of Least Squares, a best fit to any desired response may be achieved.

The MLS minimizes the sum of the squared differences of the resulting curve to the expected curve. As shown in equation 5.1, V_{ref} is a polynomial that is dependent on temperature. If the targeted curve is a constant, C, over temperature, then the MLS is minimizing the error, q, under the condition that [67]

$$q = \sum_{t=T_{\min}}^{t=T_{\max}} \left(C - V_{ref}(t)\right)^2 \quad (5.3)$$

where $T_{\max}$ is the maximum temperature and $T_{\min}$ is the minimum temperature in T_{range}. The error, q, is dependent on the scaling coefficients in the polynomial V_{ref}. The error is

minimized by taking the derivative of q with respect to each individual coefficient and setting the result equal to zero. Once this is complete, all of the coefficients may be solved for simultaneously. In the case of V_{ref}, this will result in $N+2$ equations for the $N+2$ scaling coefficients. The MLS is implemented in the following manner:

$$\begin{aligned}
\frac{\partial q}{\partial K_1} &= -2\sum_{t=T_{\min}}^{t=T_{\max}} I_{ctat}\left[C-\left(K_1 I_{ctat}+K_2 I_{ptat}-K_3 I_{PWL,1}-\cdots-K_{N+2} I_{PWL,N}\right)\right]=0 \\
\frac{\partial q}{\partial K_2} &= -2\sum_{t=T_{\min}}^{t=T_{\max}} I_{ptat}\left[C-\left(K_1 I_{ctat}+K_2 I_{ptat}-K_3 I_{PWL,1}-\cdots-K_{N+2} I_{PWL,N}\right)\right]=0 \\
\frac{\partial q}{\partial K_3} &= -2\sum_{t=T_{\min}}^{t=T_{\max}} I_{PWL,1}\left[C-\left(K_1 I_{ctat}+K_2 I_{ptat}-K_3 I_{PWL,1}-\cdots-K_{N+2} I_{PWL,N}\right)\right]=0 \qquad (5.4) \\
&\vdots \\
\frac{\partial q}{\partial K_{N+2}} &= -2\sum_{t=T_{\min}}^{t=T_{\max}} I_{PWL,N}\left[C-\left(K_1 I_{ctat}+K_2 I_{ptat}-K_3 I_{PWL,1}-\cdots-K_{N+2} I_{PWL,N}\right)\right]=0
\end{aligned}$$

In matrix form, the above relationship looks like

$$\mathbf{AK}=\mathbf{C}, \qquad (5.5)$$

where

$$\mathbf{A}=\begin{bmatrix}
-2\sum I_{ctat}I_{ctat} & -2\sum I_{ctat}I_{ptat} & -2\sum I_{ctat}I_{PWL,1} & \cdots & -2\sum I_{ctat}I_{PWL,N+2} \\
-2\sum I_{ptat}I_{ctat} & -2\sum I_{ptat}I_{ptat} & -2\sum I_{ptat}I_{PWL,1} & \cdots & -2\sum I_{ptat}I_{PWL,N+2} \\
-2\sum I_{PWL,1}I_{ctat} & -2\sum I_{PWL,1}I_{ptat} & -2\sum I_{PWL,1}I_{PWL,1} & \cdots & -2\sum I_{PWL,1}I_{PWL,N+2} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
-2\sum I_{PWL,N+2}I_{ctat} & -2\sum I_{PWL,N+2}I_{ptat} & -2\sum I_{PWL,N+2}I_{PWL,1} & \cdots & -2\sum I_{PWL,N+2}I_{PWL,N+2}
\end{bmatrix}$$

, $\mathbf{K}=\begin{bmatrix} K_1 \\ K_2 \\ K_3 \\ \vdots \\ K_{N+2} \end{bmatrix}$, and $\mathbf{C}=\begin{bmatrix} C \\ C \\ C \\ \vdots \\ C \end{bmatrix}$.

Using a mathematical tool, such as Matlab, the scaling coefficients are easily found. An example of applying the MLS to the architecture shown in Figure 5.1 will demonstrate the effectiveness of this methodology.

5.4 METHOD OF LEAST SQUARES VOLTAGE REFERENCE METHODOLOGY

A Matlab script was written to solve for the coefficients for first order, third order, and fifth order references. The script used equations 4.10 and 4.27 to generate the I_{ctat} and I_{ptat} currents. The process dependent parameters used in these equations are the same values that were used in Chapter 3 to discuss the theoretical temperature dependence of MOSFETs. The resulting temperature dependent currents I_{ctat} and I_{ptat} are shown in Figure 5.2. Using the CTAT and PTAT data, a PWL response was generated based on given temperature breakpoints. For convenience, and consistency with the examples to follow, the following relationships are defined:

$$I_{ctat} = F(t); \tag{5.6}$$

$$I_{ptat} = G(t); \tag{5.7}$$

$$I_{PWL,1} = \begin{cases} P_1F(t) - Q_1G(t) & \text{when } P_1F(t) - Q_1G(t) > 0 \\ 0 & \text{when } P_1F(t) - Q_1G(t) < 0 \end{cases}; \tag{5.8}$$

$$I_{PWL,2} = \begin{cases} B_1G(t) - D_1F(t) & \text{when } B_1G(t) - D_1F(t) > 0 \\ 0 & \text{when } B_1G(t) - D_1F(t) < 0 \end{cases}; \tag{5.9}$$

$$I_{PWL,3} = \begin{cases} P_2F(t) - Q_2G(t) & \text{when } P_2F(t) - Q_2G(t) > 0 \\ 0 & \text{when } P_2F(t) - Q_2G(t) < 0 \end{cases}; \tag{5.10}$$

$$I_{PWL,4} = \begin{cases} B_2G(t) - D_2F(t) & \text{when } B_2G(t) - D_2F(t) > 0 \\ 0 & \text{when } B_2G(t) - D_2F(t) < 0 \end{cases}, \tag{5.11}$$

where t is temperature. As described in Chapter 4, the ratios of $P:Q$ and $B:D$ set the temperature breakpoints, T_{break}, of the PWL functions. The type of PWL circuit was chosen based on the knowledge of the core's temperature response which will become apparent further in this chapter.

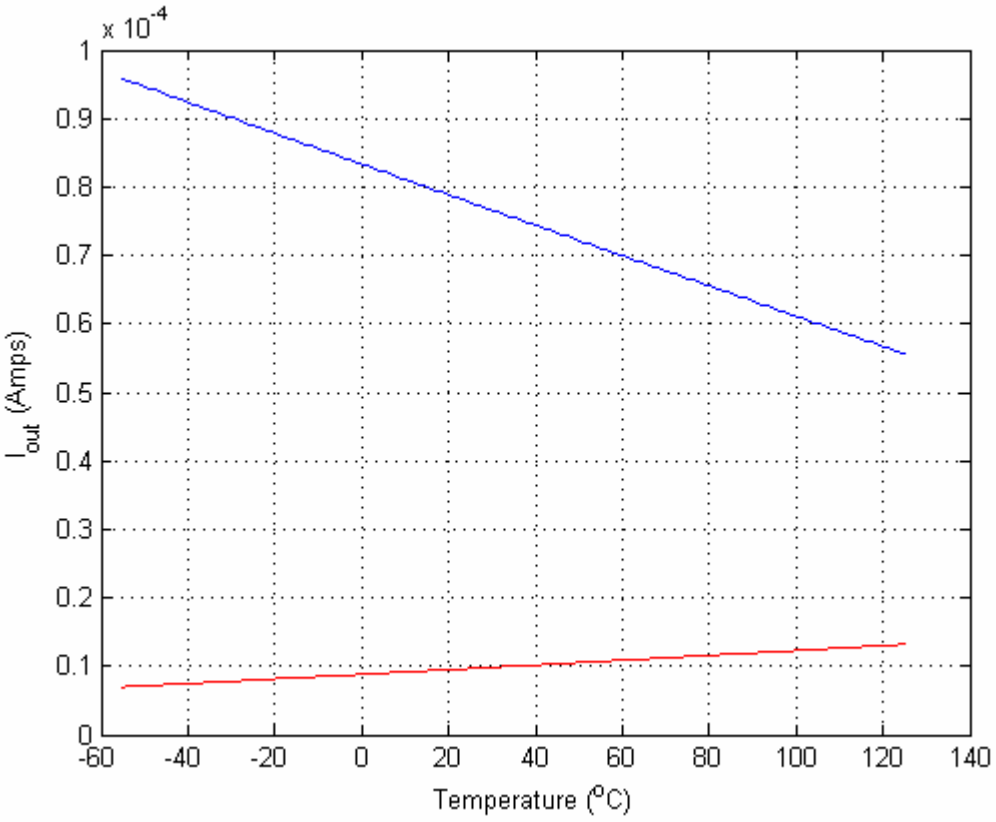

Figure 5.2: I_{ctat} and I_{ptat} for MLS Example

In the following examples, only the currents will be discussed because it is assumed that R is a scalar without any temperature dependence. This will allow a clear explanation of how the reference architecture operates and how the MLS may be applied to generate an optimum solution. The temperature range used in demonstrating the MLS method is from $-55^{o}C$ to $125^{o}C$.

The first order reference sums the scaled version of the PTAT and CTAT currents. The reference output voltage is expressed as

$$V_{ref} = \left(K_1 I_{ctat} + K_2 I_{ptat}\right)R, \tag{5.12}$$

but as mentioned earlier, the focus is on I_{total} which is expressed as

$$I_{total} = K_1 I_{ctat} + K_2 I_{ptat}. \tag{5.13}$$

Using MLS, the matrix is set up as

$$\begin{bmatrix} -2\sum I_{ctat} I_{ctat} & -2\sum I_{ctat} I_{ptat} \\ -2\sum I_{ptat} I_{ctat} & -2\sum I_{ptat} I_{ptat} \end{bmatrix} \cdot \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} 0.4/R \\ 0.4/R \end{bmatrix}. \tag{5.14}$$

The targeted output voltage is 0.4V, over the entire temperature range. For simplicity, the matrix relationship was setup as a current relationship instead of a voltage relationship for the reasons mentioned above. Solving for the scaling coefficients results in $K_1 = 0.633$ and $K_2 = 4.086$. The total current for the first order voltage reference is

$$I_{total} = 0.633 I_{ctat} + 4.086 I_{ptat}. \tag{5.15}$$

The current I_{total} is shown in Figure 5.3 and results in a parabolic curve. By estimating the maximum and minimum current over the 180 degree temperature range, the TC is approximately 30 ppm per degree Celsius. With the upward bend of the parabolic response, it is fairly straightforward to determine which type of PWL blocks will help reduce the temperature variation.

A third order voltage reference uses two PWL blocks in addition to I_{core}. The expression for the current is

$$I_{total} = K_1 I_{ctat} + K_2 I_{ptat} - K_3 I_{PWL,1} - K_4 I_{PWL,2}. \tag{5.16}$$

The PWL temperature breakpoints are set to be equally spaced apart, regardless of whether or not the I_{core} minimum is centered in the temperature range. In this case, T_{break}

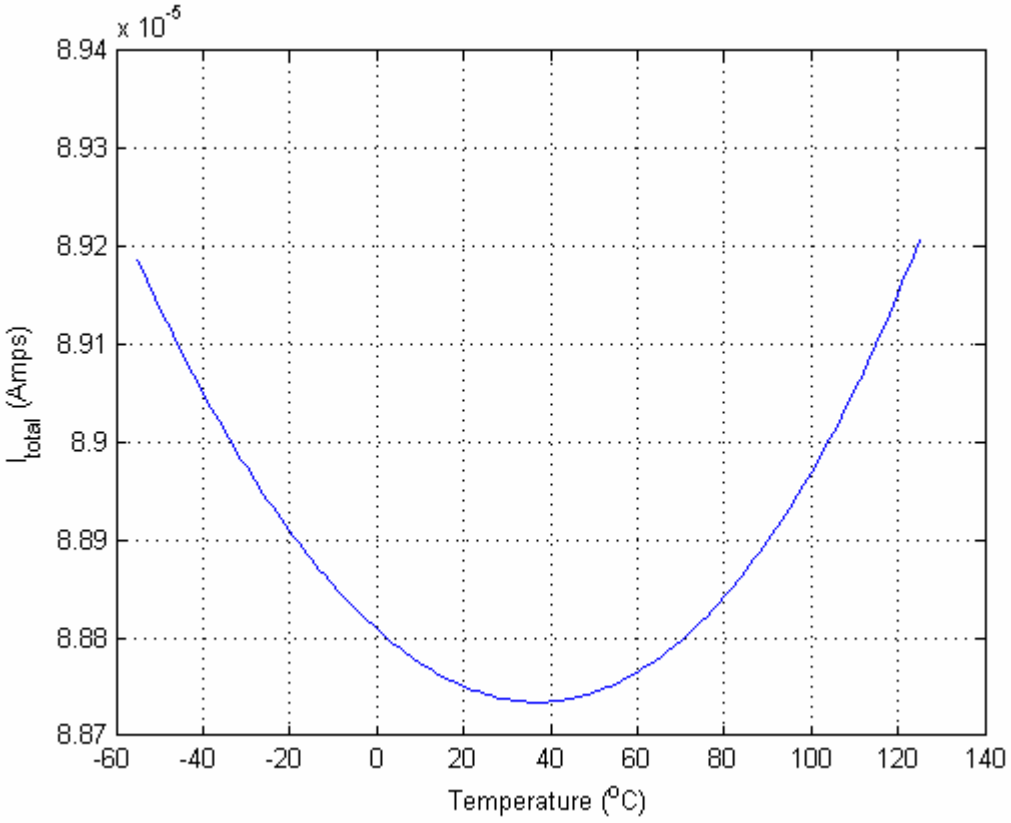

Figure 5.3: $K_1 I_{ctat} + K_2 I_{ptat}$ MLS First Order Example

of $I_{PWL,1} = 5^o C$ and T_{break} of $I_{PWL,2} = 65^o C$. With all of the currents known over temperature, the MLS may be used to find the scaling coefficients. The coefficients for the third order voltage reference are $K_1 = 0.6334$, $K_2 = 4.096$, $K_3 = 0.1094$, and $K_4 = 0.0165$. A plot of the first order response, level shifted so that its minimum is zero amps (i.e. $K_1 I_{ctat} + K_2 I_{ptat} - \min(K_1 I_{ctat} + K_2 I_{ptat})$), along with $K_3 I_{PWL,1}$ and $K_4 I_{PWL,2}$ are shown in Figure 5.4. By the relationships of these three currents, it becomes apparent that when the differences are taken between these currents, there will be less current variation over the temperature range. Figure 5.5 shows the result of equation 5.20. The TC for the third order voltage reference is 3.82 ppm per degree Celcius, which is over an order of magnitude improvement over the first order voltage reference.

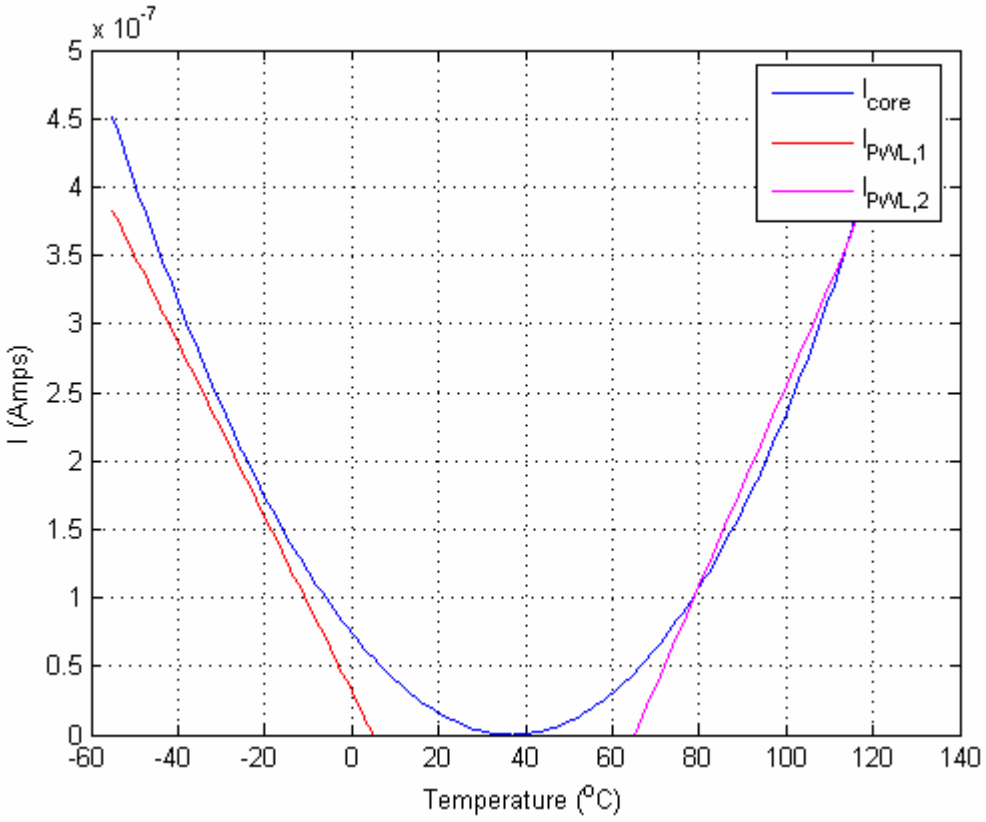

Figure 5.4: MLS Solution Plot of I_{core} Level-shifted and $K_3 I_{PWL,1}$ and $K_4 I_{PWL,2}$

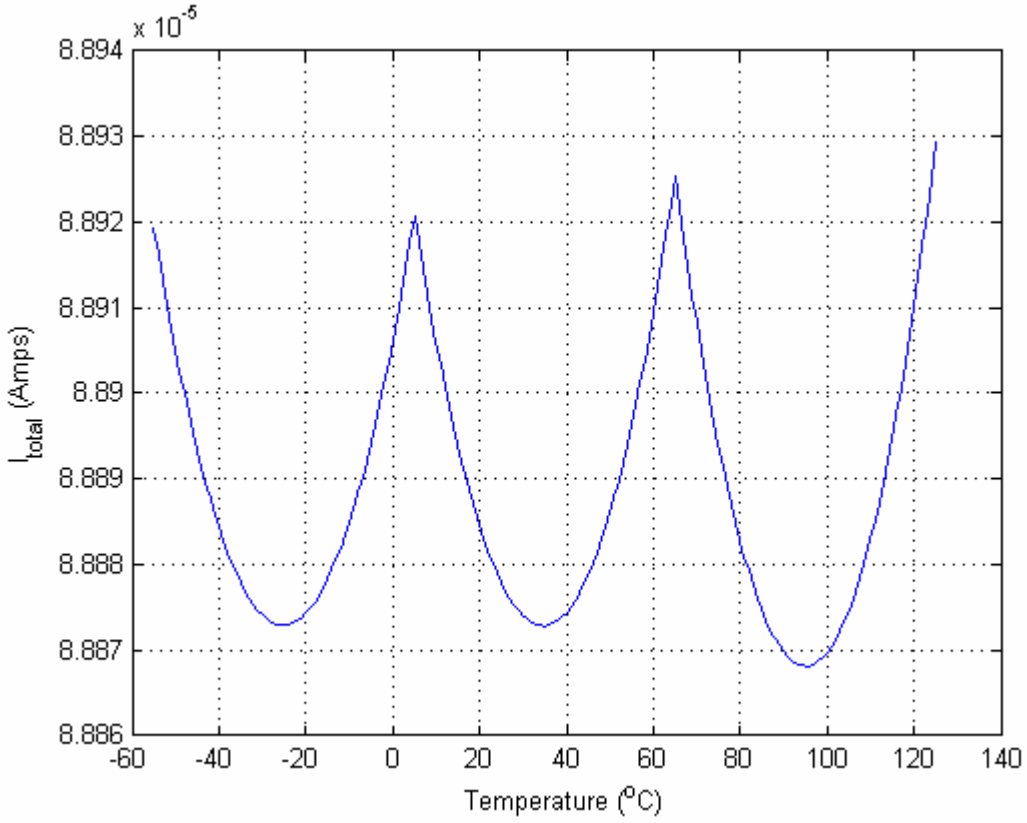

Figure 5.5: I_{total} for the Third Order Voltage Reference

The fifth order voltage reference is an extension of the third order reference. In this case, the total output current is expressed as

$$I_{total} = K_1 I_{ctat} + K_2 I_{ptat} + K_3 I_{PWL,1} + K_4 I_{PWL,2} + K_5 I_{PWL,3} + K_6 I_{PWL,4}. \qquad (5.17)$$

Again in this case, the temperature breakpoints of the PWL blocks are set apart by equal temperatures: $T_{break} = -19^o C$ for $I_{PWL,1}$; $T_{break} = 17^o C$ for $I_{PWL,2}$; $T_{break} = 53^o C$ for $I_{PWL,3}$; $T_{break} = 89^o C$ for $I_{PWL,4}$. Solving for the scaling coefficients using the MLS results in $K_1 = 0.6335$, $K_2 = 4.096$, $K_3 = 0.0695$, $K_4 = 0.0643$, $K_5 = 0.00913$, and $K_6 = 0.0114$. A plot of the level shifted I_{core} with all of the PWL currents are shown in Figure 5.6. A plot of I_{total} is shown in Figure 5.7 and the TC is 1.4 ppm, which is approximately half of the TC of the third order voltage reference.

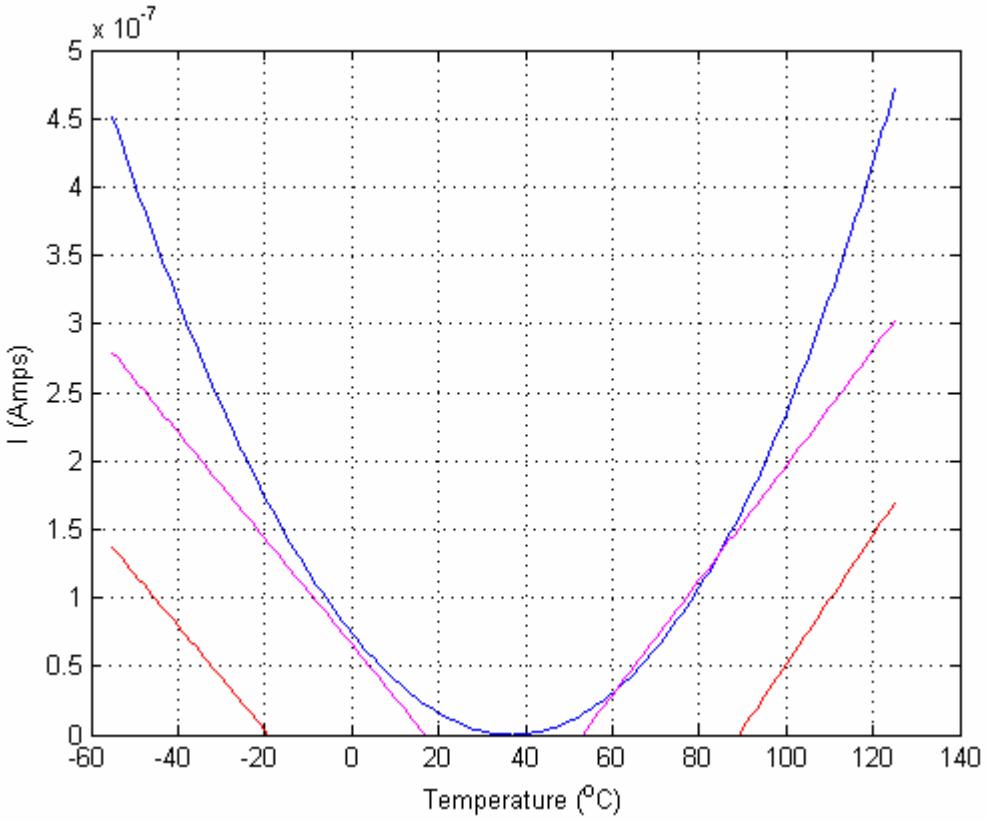

Figure 5.6: I_{core} Level-shifted with $K_3 I_{PWL,1}$, $K_4 I_{PWL,2}$, $K_5 I_{PWL,3}$, and $K_6 I_{PWL,4}$

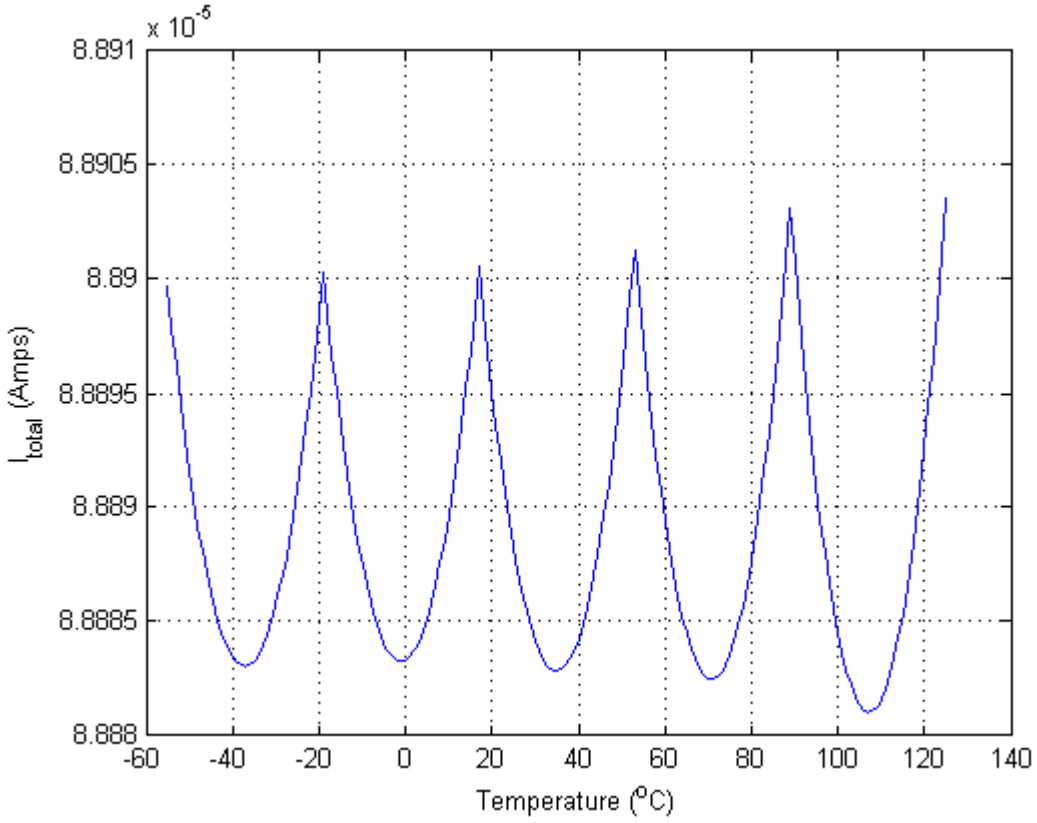

Figure 5.7: I_{total} for the Fifth Order Voltage Reference

5.5 OPTIMIZING TC

The MLS minimizes the square of the integrated error over the temperature range. By inspecting Figures 5.5 and 5.7, it should be apparent that the integrated error over the temperature range was minimized but the TC was not. The MLS minimized the error based on the temperature breakpoints that the algorithm was given. In order to find the optimized solution for a particular higher order voltage reference, proper selection of the PWL temperature breakpoints is required. As mentioned above, the MLS was performed using a Matlab script. By expanding this script to do an iterative search for the best measured TC by sweeping the PWL temperature breakpoints, the optimum

solution may be found. Figure 5.8 shows the results of a third order voltage reference where the breakpoints are set and where the optimum TC was found by sweeping the PWL temperature breakpoints.

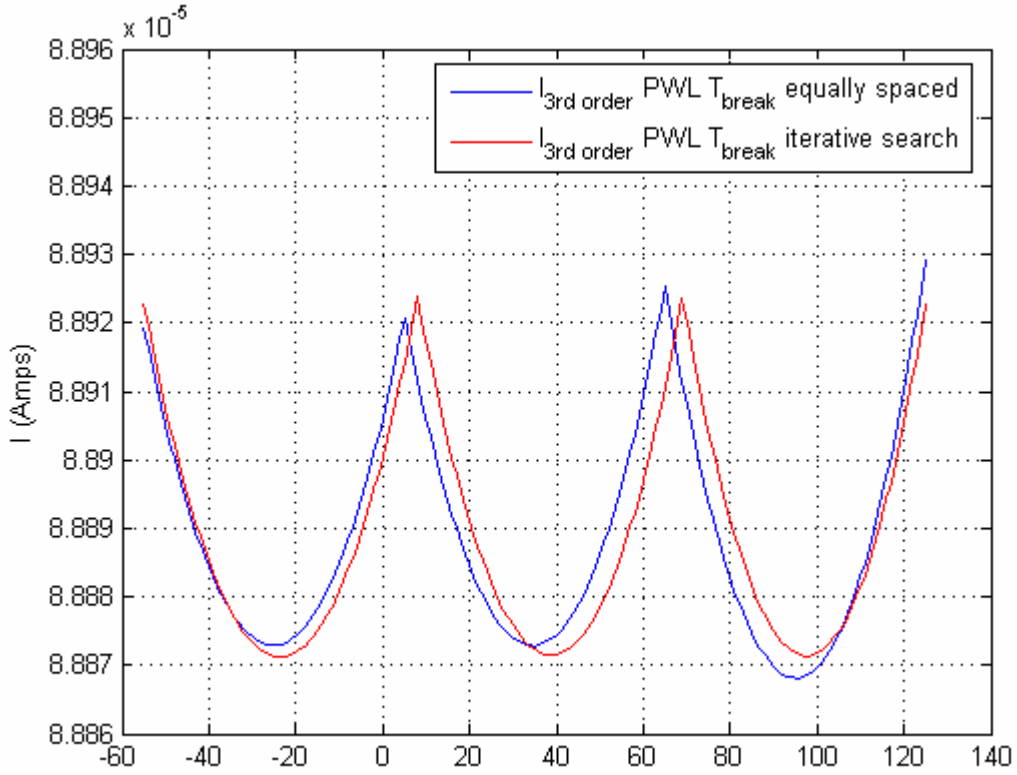

Figure 5.8: I_{total} of Set and Iteratively Searched PWL Temperature Breakpoints

The resulting improvement is just over 14%, bringing down the TC to 3.28 ppm per degree Celsius. The original temperature breakpoints were set at $5^{o}C$ and $65^{o}C$. When the optimum solution was sought by sweeping the temperature breakpoints, the algorithm settled on $8^{o}C$ and $69^{o}C$.

5.6 MLS METHODOLOGY CONCLUSION

A simple and expandable voltage reference architecture has been shown. By using the Method of Least Squares in combination of an iterative search for the PWL

temperature breakpoints, an optimum solution for a voltage reference may be found. The TC was reduced by an order of magnitude by increasing the voltage reference from a first order to a third order reference. By increasing the order again, to a fifth order voltage reference, the TC was reduced by an additional half. By cascading N number of PWL blocks to a first order reference, further reductions are possible.

Chapter 6: Voltage Reference Circuit

6.1 INTRODUCTION

This chapter uses the temperature dependent circuits that were developed in Chapter 4 and applies the design methodology discussed in Chapter 5 to produce stable voltage references. First order, third order, and fifth order references are first presented in a 0.18μ process technology operating with a power supply of 0.9 volts and producing a reference voltage of 0.4 volts. Then, first and third order voltage references are presented in a $90nm$ process technology operating at 0.4 volts with an output voltage of 0.2 volts. In both technologies, the power supply was targeted to be approximately twice that of the threshold voltages of the MOSFET devices. The impact of process variations and techniques for calibrating or trimming the voltage references are discussed.

6.2 FIRST ORDER VOLTAGE REFERENCE, 0.18μ

The summation of a PTAT and CTAT current into a resistor forms the core of the voltage reference. It is from this first order voltage reference that higher order voltage references will be implemented using the PWL circuits. Figure 6.1 shows a first order reference, based on one of the temperature dependent circuits that were discussed in Chapter 4.

The box labeled "I_{ptat} and V_{ctat} generator" was described in detail and may be reviewed in section 4.3 where Figure 4.6 is presented. As previously presented, the

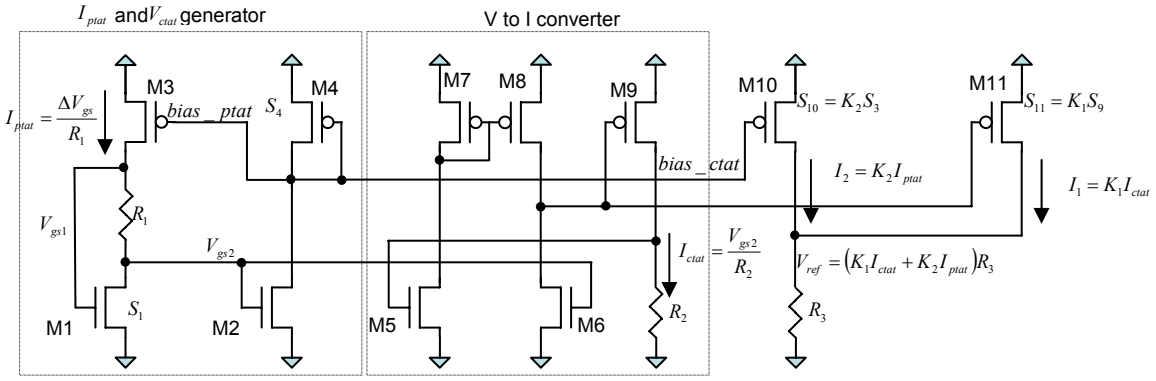

Figure 6.1: Voltage Reference Core

current I_{ptat} is generated through device M3 based on R_1 and the ΔV_{gs} of devices M1 and M2 so that

$$I_{ptat} = \frac{\Delta V_{gs}}{R_1}. \tag{6.1}$$

This circuit is a self-sustaining circuit. In other words, it's a closed loop feedback system that holds itself at a desired operating point. There is an inherent startup problem with this circuit that is easily resolved with an additional circuit. The additional circuit should ensure that the voltage reference turns on while not interfering with the performance of the voltage reference when it is operational. There are many published examples of startup circuits that are published [15, 69, 74, 85].

A voltage to current converter, labeled "V to I converter," produces I_{ctat} that is proportional to the V_{gs2} voltage. This circuit is comprised of an OPAMP that drives a current to force the voltage across R_2 to be equal to V_{gs2}. Therefore,

$$I_{ctat} = \frac{V_{gs2}}{R_2}, \tag{6.2}$$

where I_{ctat} is being driven by device M9. The amount of error of this current is related to the open loop gain of the OPAMP. The OPAMP's quiescent current is regulated by the ratio of M2 to M6. A stability problem may exist with the OPAMP, but stabilization

techniques are well known and may be applied here [7, 15, 33, 74]. Typically, a capacitor is used and if one is required, it may be placed across the gate and drain of device M5.

The temperature dependent currents through devices M3 and M9 will be used by the MLS algorithm that was presented in Chapter 5 to set the scaling coefficients. The voltage reference scales I_{ctat} and I_{ptat} by K_1 and K_2 by setting the ratio of the current density of devices M11 to M9 and M10 to M3, respectfully. As a result the reference voltage output may be expressed as

$$V_{ref} = \left(K_1 \frac{V_{gs2}}{R_2} + K_2 \frac{\Delta V_{gs}}{R_1} \right) R_3 . \tag{6.3}$$

It is assumed that the material is the same for all of the resistors. Thus, the temperature variation of resistors R_1 and R_2 are cancelled out by the temperature variation of R_3. To be consistent with Chapter 5, it is worth noting that the total current for this circuit was defined as I_{total} and may be expressed as

$$I_{total} = K_1 \frac{V_{gs2}}{R_2} + K_2 \frac{\Delta V_{gs}}{R_1} . \tag{6.4}$$

In order to merge the MLS algorithm with the circuit design, a Spice simulation was run on the first order voltage reference to collect the temperature dependent currents from devices M3 and M9. Using the same algorithm that was used in Chapter 5, but replacing the theoretical data with the Spice simulated I_{ctat} and I_{ptat} data, the scaling coefficients were determined. This has the benefit of including the non-ideal effects of the temperature dependent circuits.

The MLS algorithm does not consider the non-ideal effects of transistors M10 and M11. Therefore, a slight adjustment of device ratios is required in order to minimize the TC. The adjustment that was needed in this case moved the minimum of the parabolic temperature response only a few degrees. The temperature response for

$0^o C < T < 70^o C$ and $-55^o C < T < 125^o C$ are shown in Figure 6.2 and Figure 6.3, respectfully. The TC of these two temperature sweeps are 56 $ppm/^o C$ for the $0^o C < T < 70^o C$ case and 132 $ppm/^o C$ for the $-55^o C < T < 125^o C$ case.

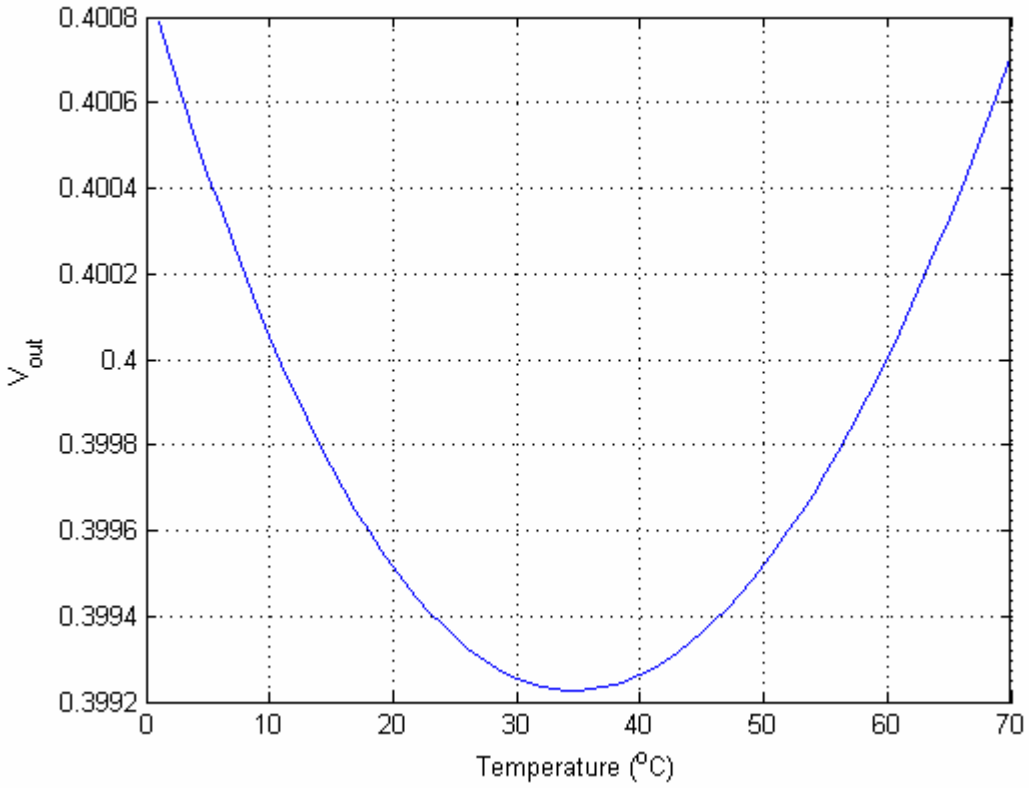

Figure 6.2: 0.18μ First Order Reference for $0^o C < T < 70^o C$

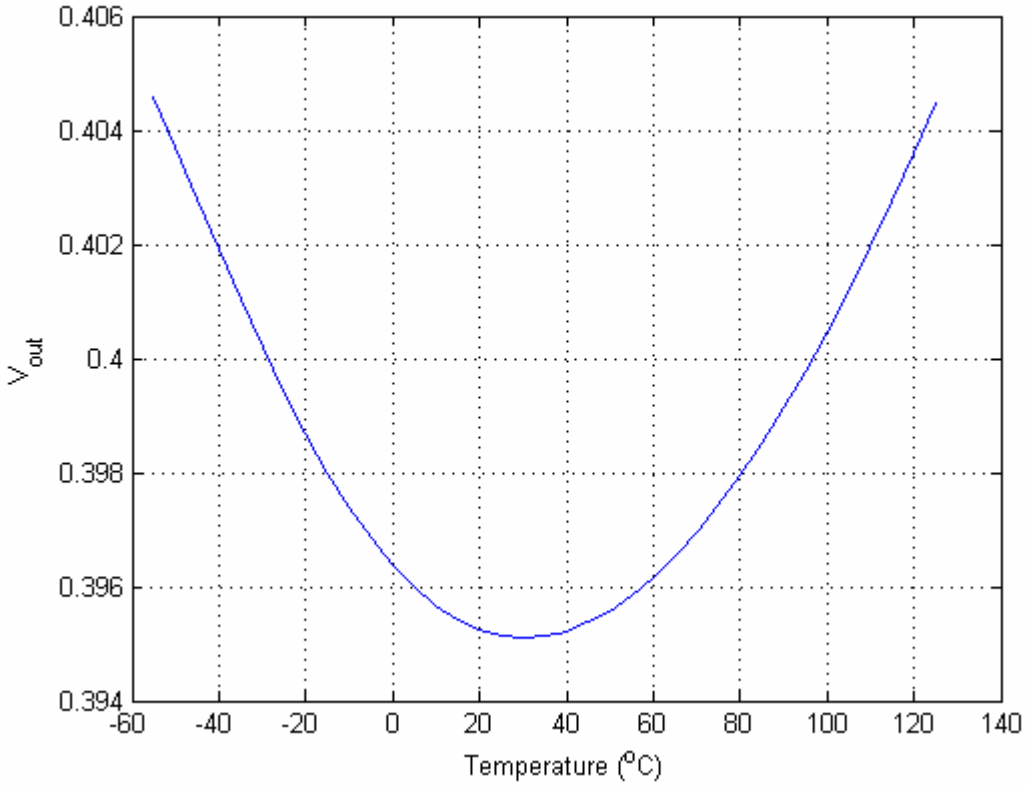

Figure 6.3: 0.18μ First Order Reference for $-55^{o}C < T < 125^{o}C$

6.3 THIRD ORDER VOLTAGE REFERENCE, 0.18μ

A third order voltage reference is created by adding two PWL circuits to the circuit presented in Figure 6.1. Figure 6.4 shows the additional PWL circuits. It is worth noting that these PWL circuits are mathematically equivalent to the ones presented in section 4.4. However, instead of sourcing a current, these circuits will sink the desired current to correct for the parabolic curvature observed in Figures 6.2 and 6.3.

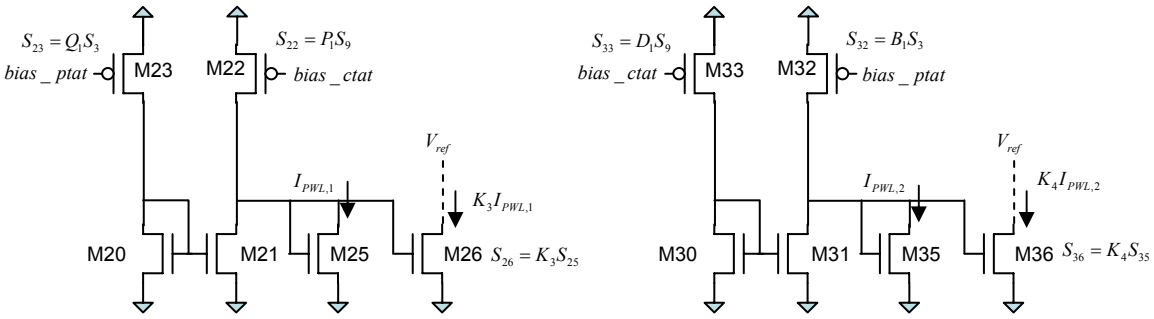

Figure 6.4: PWL Circuits for Third Order Voltage Reference

The MLS algorithm uses the simulated I_{ctat} and I_{ptat} data while sweeping the temperature breakpoints of the PWL circuits. Similar to the first order voltage reference, the non-ideal effects of the transistors require some final adjustments to the scaling coefficients. The ratios that the MLS algorithm chose for $B:D$ and $P:Q$ set the temperature breakpoints. There were no necessary adjustments to these ratios, but there were some adjustments required for the scaling coefficients K_3 and K_4. The results of the third order voltage reference design for the $0^o C < T < 70^o C$ and $-55^o C < T < 125^o C$ case are shown in Figures 6.5 and 6.6, respectively. The performance results of these two cases are $4.3\, ppm/^o C$ for $0^o C < T < 70^o C$ and $11.3\, ppm/^o C$ for $-55^o C < T < 125^o C$. It is shown that temperature variations are reduced by approximately an order of magnitude going from the first order to the third order voltage reference. These results are consistent with the findings and conclusions of the system design and analysis in Chapter 5.

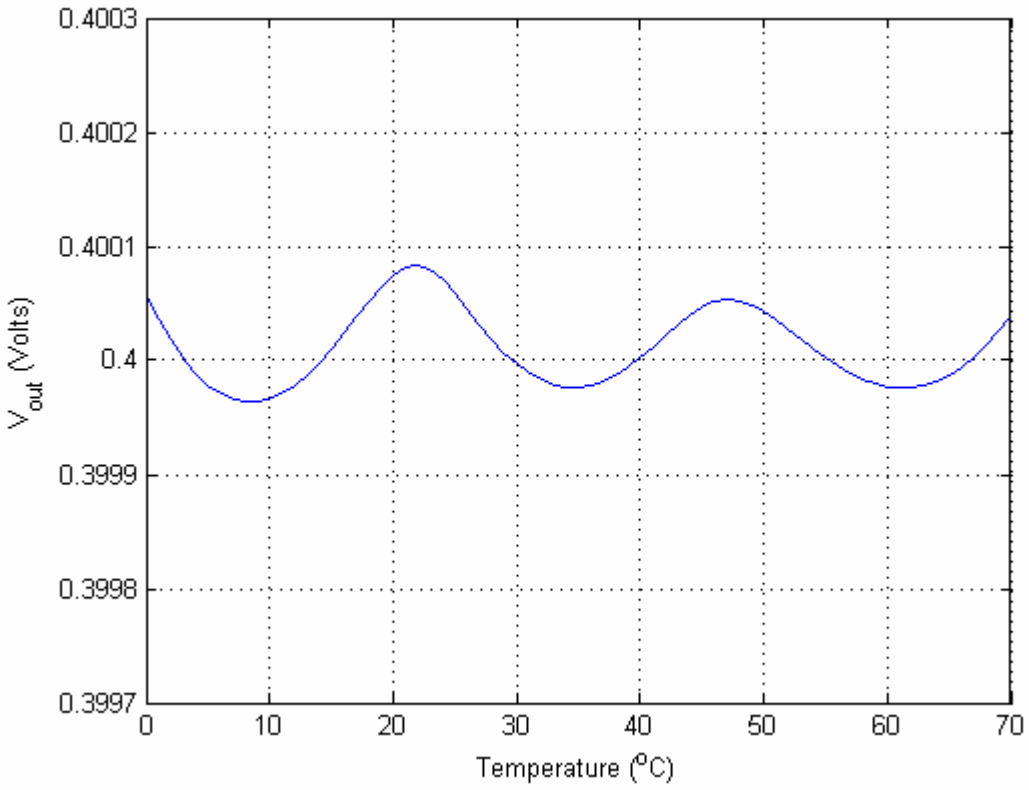

Figure 6.5: 0.18μ Third Order Reference for $0^{o}C < T < 70^{o}C$

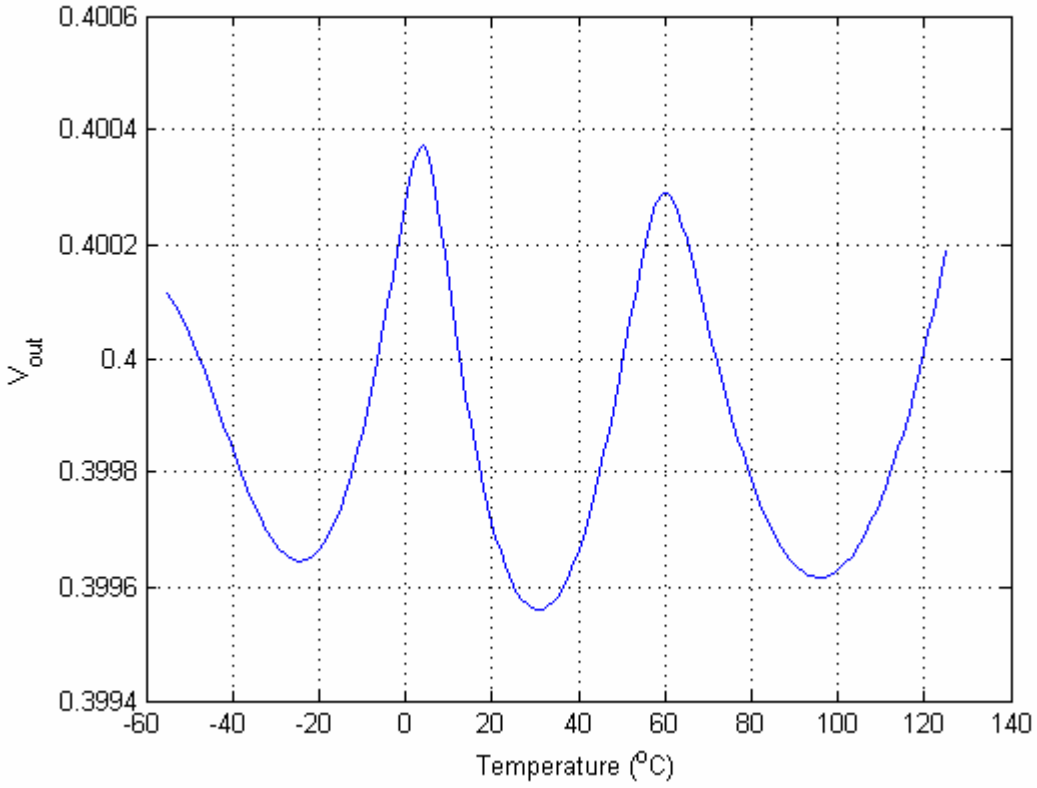

Figure 6.6: 0.18μ Third Order Reference for $-55^{o}C < T < 125^{o}C$

6.4 FIFTH ORDER VOLTAGE REFERENCE, 0.18μ

A fifth order voltage reference is an extension of the third order voltage reference. Adding an additional circuit, as presented in Figure 6.4, results in a total of 4 PWL circuits. Two of the PWL circuits subtract the currents from the parabolic curve before the minimum. The other two PWL circuits subtract currents when the temperature is above the minimum.

The MLS algorithm is set up as it was used for the scaling coefficients in the third order voltage reference example. The non-ideal effects of the transistors had some effects that were not predicted by the MLS algorithm, so again, some minor adjustments

were required. The results of the fifth order voltage reference are shown in Figure 6.7 for the $-55^{o}C < T < 125^{o}C$ case. The performance in this case is $4.8\,ppm/^{o}C$, which is twice the TC performance of the third order voltage reference. This is consistent with the system analysis of Chapter 5.

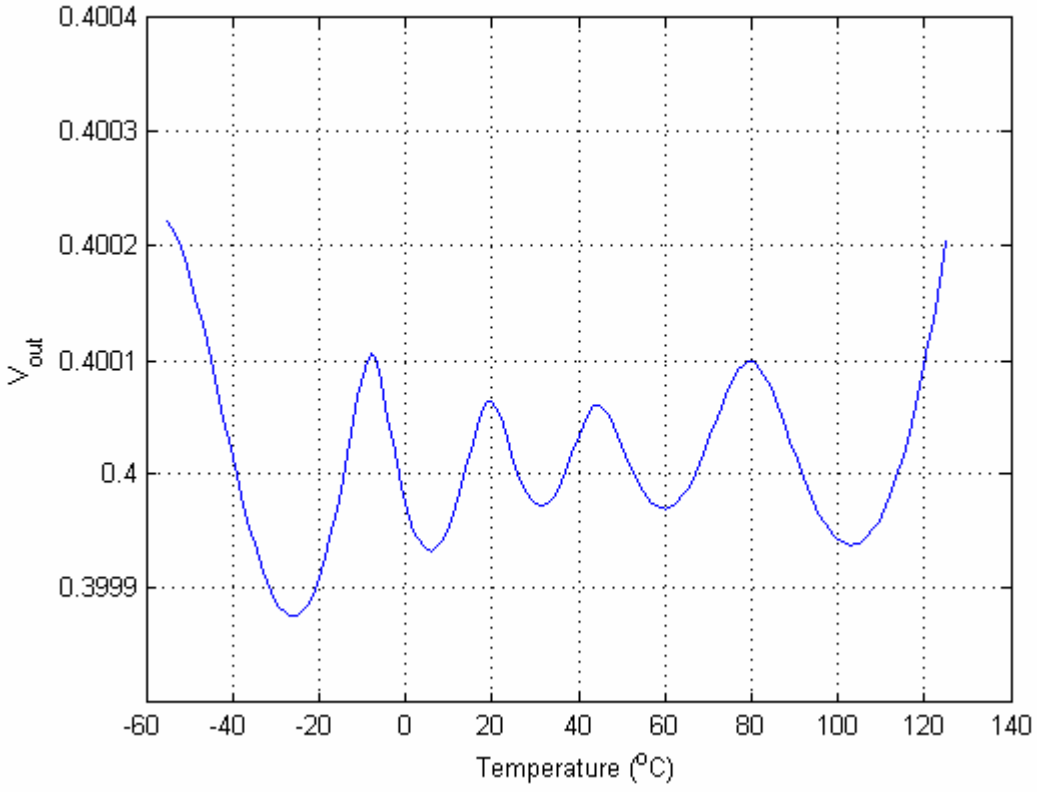

Figure 6.7: 0.18μ Fifth Order Reference for $-55^{o}C < T < 125^{o}C$

6.5 FIRST ORDER VOLTAGE REFERENCE, 0.09μ

The first order voltage reference circuit used for the $90nm$ process may be observed in Figure 6.1. As expected, this architecture is fully capable of being scaled with technology. As discussed in Chapter 2, some of the non-ideal effects of the transistors are more apparent in this process. For example, the leakage currents of the devices become apparent at the higher temperatures when the same magnitudes of the

I_{ctat} and I_{ptat} currents are used in the 0.18μ design example. By decreasing the resistor sizes and increasing the I_{ctat} and I_{ptat} currents, it is possible to reduce the relative contributions of the leakage currents. This increases the overall power consumption, but is required to effectively swamp out as much of the leakage current as possible.

Using the simulated data for I_{ctat} and I_{ptat}, the MLS algorithm solves for the scaling coefficients. A plot of the first order voltage reference response to temperature is shown in Figure 6.8 and Figure 6.9 for the $0^oC < T < 70^oC$ and $-55^oC < T < 125^oC$ case, respectively. The TC performance of the voltage reference is $108\,ppm/^oC$ for $0^oC < T < 70^oC$ and $230\,ppm/^oC$ for $-55^oC < T < 125^oC$. The temperature variation for the $90nm$ process is about twice that of the 0.18μ process. This can be attributed to the leakage current increasing with temperature. It can be observed that the minimum of the response curve is no longer centered at 35^oC, but is offset to a higher temperature. The leakage current from devices M10 and M11 are contributing to I_{ptat}, thus effectively changing the scaling coefficient K1 with temperature. If it is desired to compensate for this effect, a p-channel dummy load may be added. The dummy device should be configured so that the drain, gate, and bulk are tied to V_{ref} and the source is tied to V_{ss}. The appropriate size device would allow the dummy load's leakage current to source the leakage current of M10 and M11, effectively canceling those currents from the summing sources.

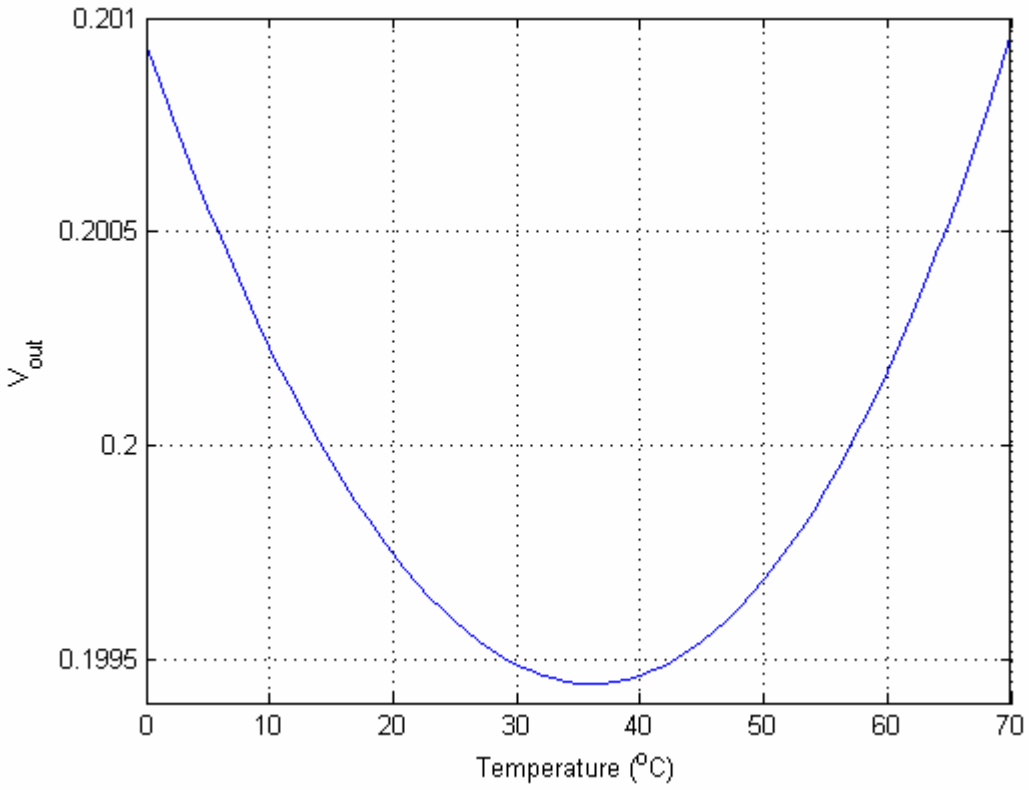

Figure 6.8: $90nm$ First Order Reference for $0^{o}C < T < 70^{o}C$

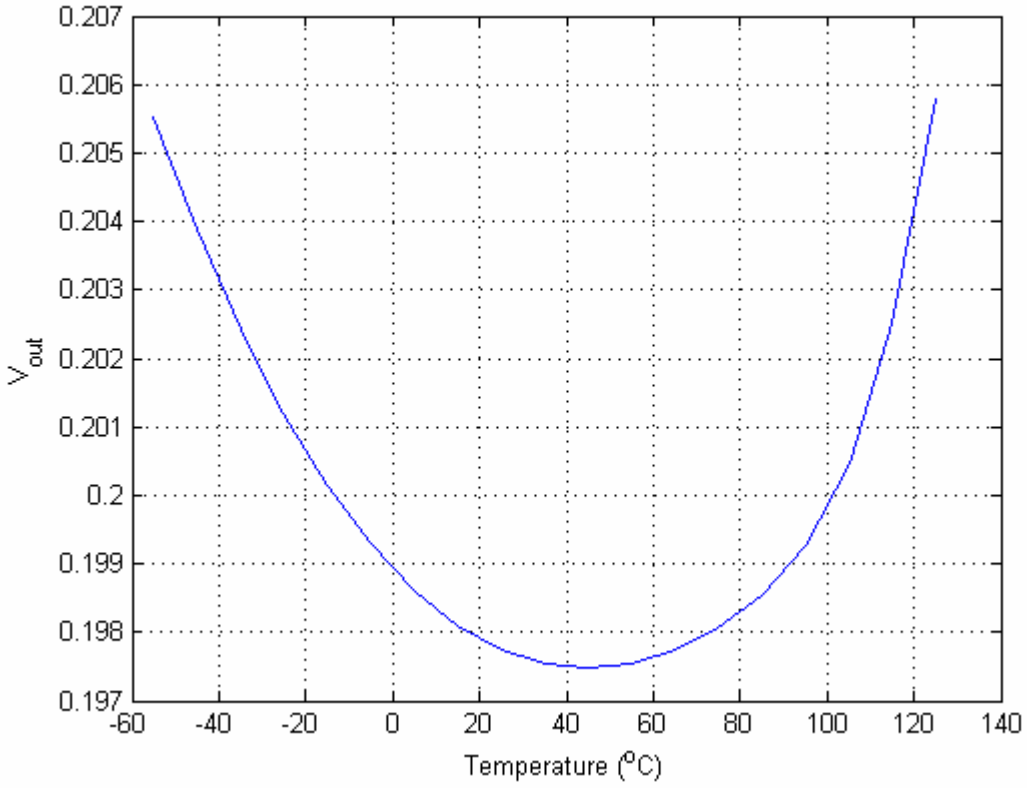

Figure 6.9: $90nm$ First Order Reference for $-55^{o}C < T < 125^{o}C$

6.6 THIRD ORDER VOLTAGE REFERENCE, 0.09μ

The third order voltage reference for the $90nm$ process is implemented by combining the circuits shown in Figures 6.1 and 6.4. Due to the non-ideal effects of the transistors in this process, a modification was done to the PWL circuit that sinks the current $I_{PWL,2}$. Due to the output impedance of the devices and their leakage currents, there was not a strong transition as the temperature varies across the temperature breakpoint. Instead, the transition is rounded and difficult to use in the final solution of the circuit.

A differential approach was taken to solve this problem. Figure 6.10 shows a more complex circuit that accomplishes the PWL action that is desired. This circuit is comprised of two PWL circuits that are tied to the input of a two stage differential OPAMP. The two PWL circuits are designed so that one is active when the temperature is below a specified temperature breakpoint. The other PWL circuit is designed to be active above the same specified temperature breakpoint. Thus, the two currents generated will cross each other at T_{break}, regardless of whether or not there is a strong transition from either circuit. The OPAMP is used to amplify, convert the differential signal into a single ended output, and drive the $I_{PWL,2}$ current with higher output impedance devices.

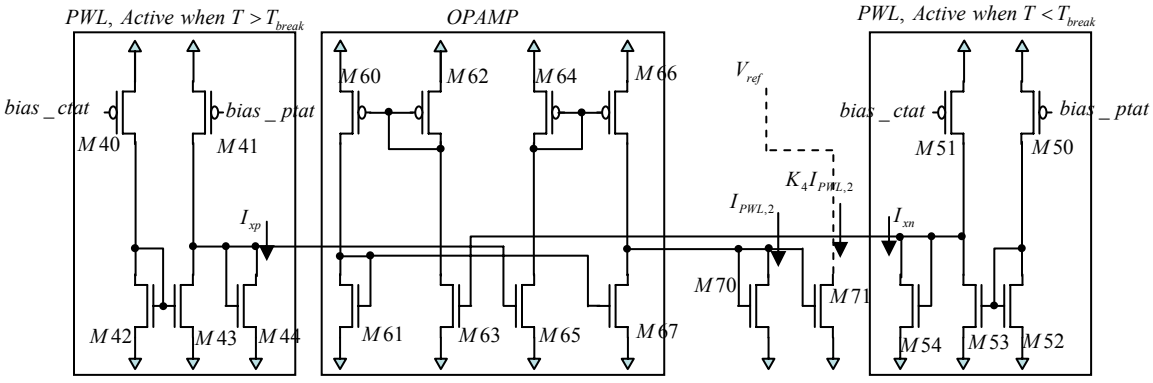

Figure 6.10: Gain Enhanced PWL Circuit

The same methodology that was used in section 6.2 is used with the MLS algorithm to solve for the scaling coefficients and temperature breakpoints. The result for the third order voltage reference is shown in Figure 6.11. The TC is measured to be $7.3\ ppm/^{o}C$, which is approximately an order of magnitude improvement over the first order voltage reference in the $90nm$ process.

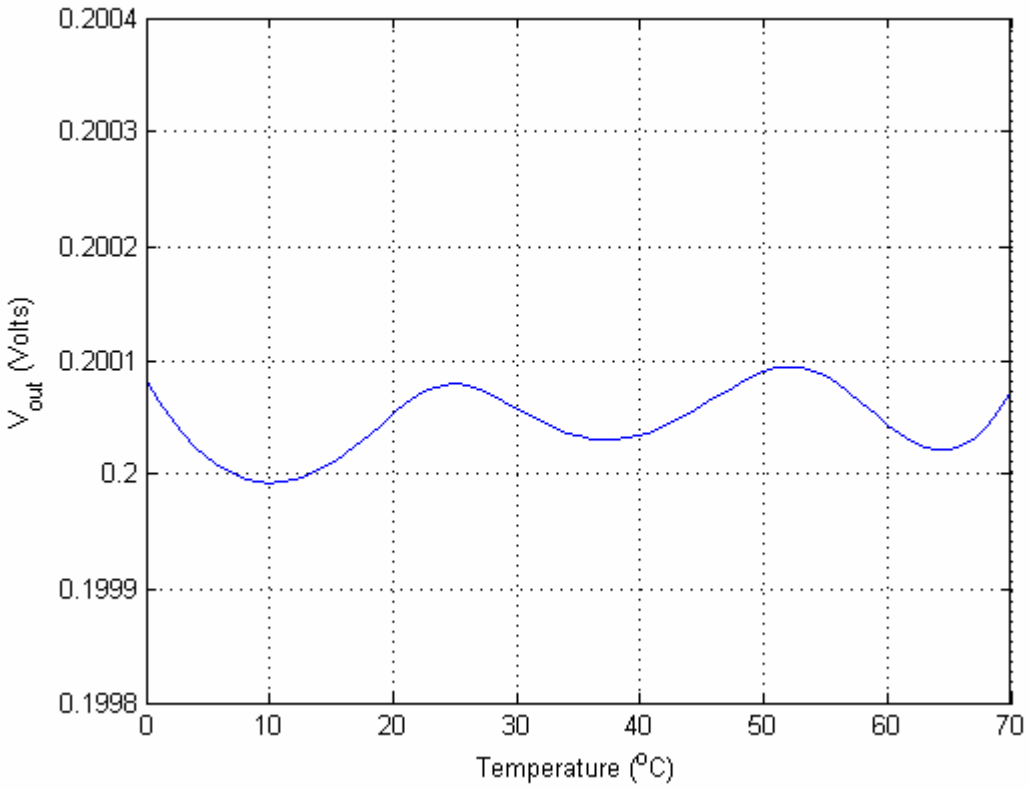

Figure 6.11: $90nm$ Third Order Reference for $0^{o}C < T < 70^{o}C$

6.7 DEPENDENCE ON POWER SUPPLY

All circuits have some response to power supply variations. Some circuits are more immune to these variations than others. Figure 6.12 shows the results of a power supply sweep from 0.1 volts to 2.0 volts for the circuit used in the 0.18μ process. It can be observed that the circuit is active for a power supply voltage that is above 0.6 volts. This, in effect validates the capability of this architecture to operate down to extremely low voltages. Figure 6.13 shows the results of a power supply sweep from 0.1 volts to 1.0 volts for the circuit designed in the $90nm$ process. In this case, it seems that the circuit becomes active in the region where the power supply reaches 0.3 volts.

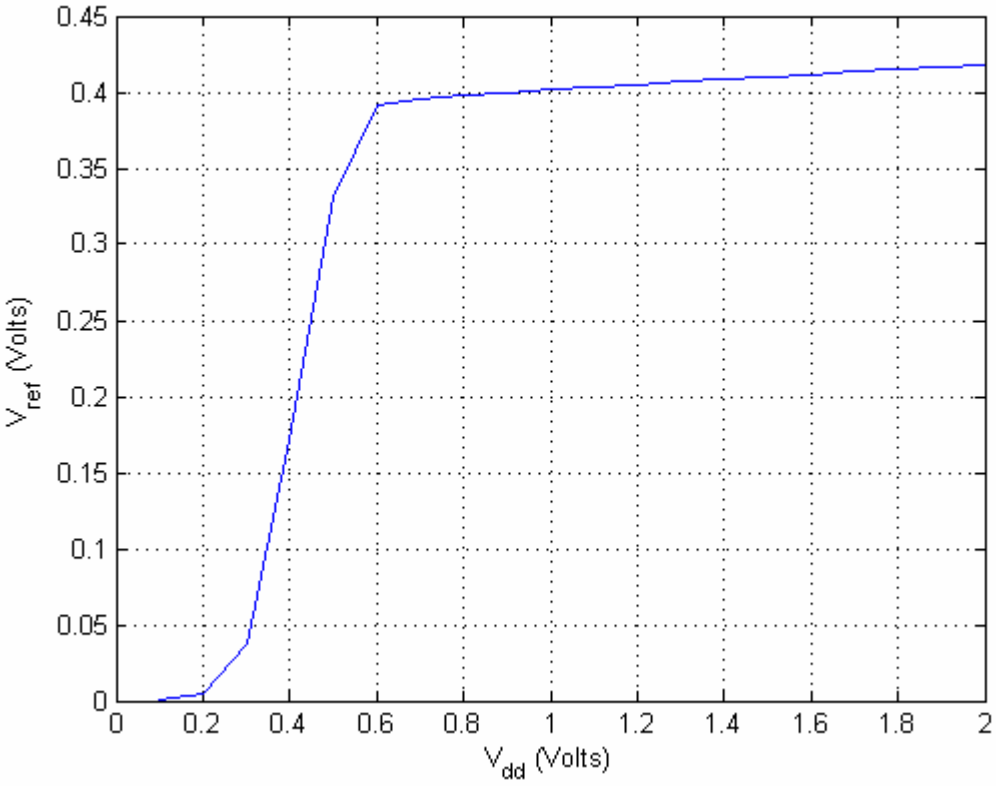

Figure 6.12: Power Supply Sweep of Voltage Reference (0.18μ)

In both cases, the response to the power supply when the circuits are fully functional seems to indicate a reasonable immunity to the power supply variation. The power supply variation is on the order of $30dB$. This can be improved significantly by swapping out the core of the voltage reference. If the circuit presented in Figure 4.5 is used, then the expected power supply rejection is expected to increase. The OPAMP in that design provides a significant increase of DC power supply rejection. A previous design using this architecture was completed and was measured to have a power supply rejection on the order of $80dB$.

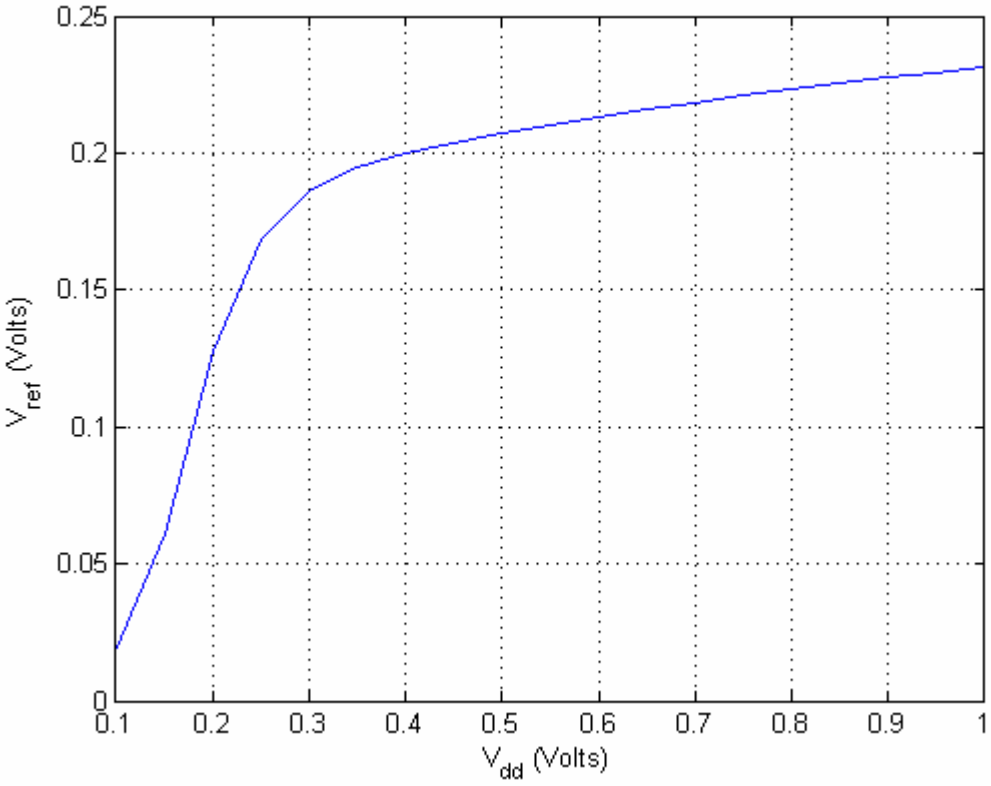

Figure 6.13: Power Supply Sweep of Voltage Reference ($90nm$)

6.8 PROCESS VARIATION

Several process parameters affect the performance of the voltage reference circuit [89-92]. This circuit architecture was setup to help minimize the errors introduced in manufacturing. Device and resistor mismatches change the operating points of the CTAT and PTAT circuits. Device mismatches cause relative errors in the scaling coefficients. All of these mismatches contribute to deviations from the ideal output voltage and the TC. The matching of device M1 to M2 sets the operating point for the temperature dependence of I_{ptat}. In the same manner, the relative operating point of I_{ctat} to I_{ptat} is dependent on the matching of resistors R_1 and R_2. A voltage offset error may

also be introduced by the OPAMP's input pair from devices M5 and M6, causing error in I_{ctat}. The scaling coefficient accuracy is dependent on matching devices M9 and M3 to all of the PWL circuits and the devices that make up the core of the voltage reference. Any process variations of these devices alter the required scaling coefficients for proper placement of temperature breakpoints as well as scaling the amount of current to sum into the output voltage's node. Depending on the application's need for accuracy, the process variation may be addressed by the physical circuit or by performing a trim or calibration to the circuit after manufacturing.

For the architecture presented in Figure 6.1, both the I_{ctat} and I_{ptat} operating points depend on the same current in the feedback loop. There are two possible opportunities for a post manufacturing adjustment for this circuit. First, a current may source from or sink into the top of R_1. By injecting a positive or negative error current into this node, the ideal operating point of this circuit may be regained. This should position the V_{ref} closer to its ideal voltage while optimizing the TC. A current DAC may be designed to accomplish this goal, where the calibration data may be stored in memory. The second method, which is frequently used for accurate voltage references today, includes laser trimming the resistors so that the voltage reference returns to its optimum operating point.

Minimizing process variation is also achieved by optimizing the area of the devices [89-92]. It can be shown that the threshold voltage of closely spaced MOSFETs will vary by the approximation of [15]

$$\sigma(V_{th}) \approx \frac{0.1t_{ox}}{\sqrt{WL}}. \tag{6.5}$$

Therefore, to decrease the threshold voltage mismatch it is desirable to make the desired matching devices as large as possible. Resistor mismatch is highly dependent on the process and the manufacturer of the silicon. A general scaling relationship of the

variation relative to the process technology is not available, but local resistor mismatch variation will scale proportionally to

$$1/\sqrt{R_{area}}\ , \tag{6.6}$$

where R_{area} is the area of a local resistor. However, each manufacturing facility will have statistical data available. For the 0.18μ process technology, it is assumed that both the n-channel and p-channel devices have a process variation of $\sigma(V_{th}) = 5mV/\sqrt{WL}$. The resistors are assumed to have a process variation of $\sigma(R) = 0.18\%$ for the width and lengths used. Both one sigma variations used here are slightly altered numbers from actual processing data reports provided by the manufacturer. This type of data is considered confidential by the manufacturers, so it can not be disclosed in this document.

Data was captured for 500 simulation runs of the first order, third order, and fifth order circuit designs in the 0.18μ process. All MOSFET devices had the threshold voltages varied independently by the relationship discussed above. Also, all of the resistors had their values independently adjusted by the variation discussed above. A histogram plot of all three designs, for the temperature range of $-55^oC < T < 125^oC$, is shown in Figure 6.14. Figure 6.15 places the statistical data of all three designs onto the same plot to emphasize the performance increase with each order of the voltage reference. A summary of the expected circuit performance in manufacturing is summarized in Table 6.1.

Process variation simulations were completed for the $90nm$ process over the temperature range of $0^oC < T < 70^oC$ for the first order and third order voltage references. Mismatch and process variation data was not available for the $90nm$ process, so this simulation is not accurate but gives a general indication of its present sensitivity to process variation. The results are shown in Figure 6.16 and Figure 6.17. The statistical results are summarized in Table 6.2. Not much emphasis was placed on the design for

process variation for this circuit, and it is apparent. The assumptions for variation are the same as that for the 0.18μ process, except that the $\sigma(V_{th})$ was set to $2.5\,mV/\sqrt{WL}$. The resistor mismatch variation remained the same. Other dominant mismatch sources start to appear in this process but were not modeled, such as the mismatch of gate leakage currents due to the gate-oxide thickness variations.

Table 6.1: 1^{st}, 3^{rd}, and 5^{th} Order Process Variations of 0.18μ, $-55^{o}C < T < 125^{o}C$

1^{st} Order	$mean(V_{ref}) = 0.3985$ $\sigma(V_{ref}) = 3.19mV$ $mean(TC) = 132\ ppm/^{o}C$ $\sigma(TC) = 0.675\ ppm/^{o}C$
3^{rd} Order	$mean(V_{ref}) = 0.3998$ $\sigma(V_{ref}) = 1.83mV$ $mean(TC) = 15.28\ ppm/^{o}C$ $\sigma(TC) = 3.55\ ppm/^{o}C$
5^{th} Order	$mean(V_{ref}) = 0.3999$ $\sigma(V_{ref}) = 1.96mV$ $mean(TC) = 10.3\ ppm/^{o}C$ $\sigma(TC) = 4.24\ ppm/^{o}C$

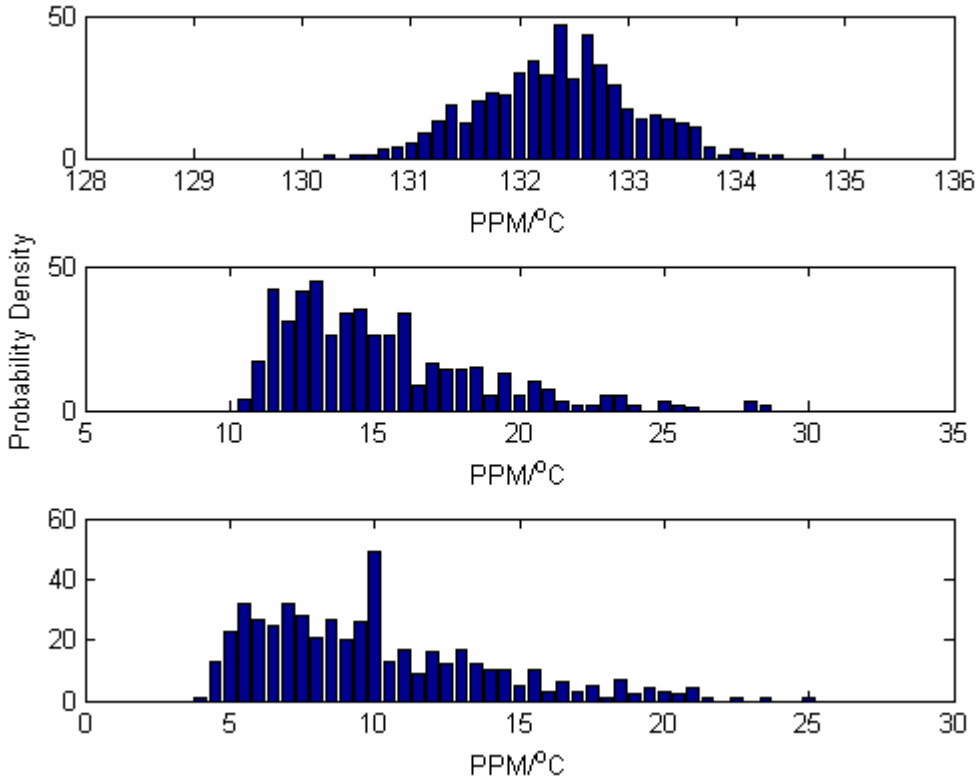

Figure 6.14: Process Variations of 1^{st}, 3^{rd}, and 5^{th} Order References, $-55^{o}C < T < 125^{o}C$

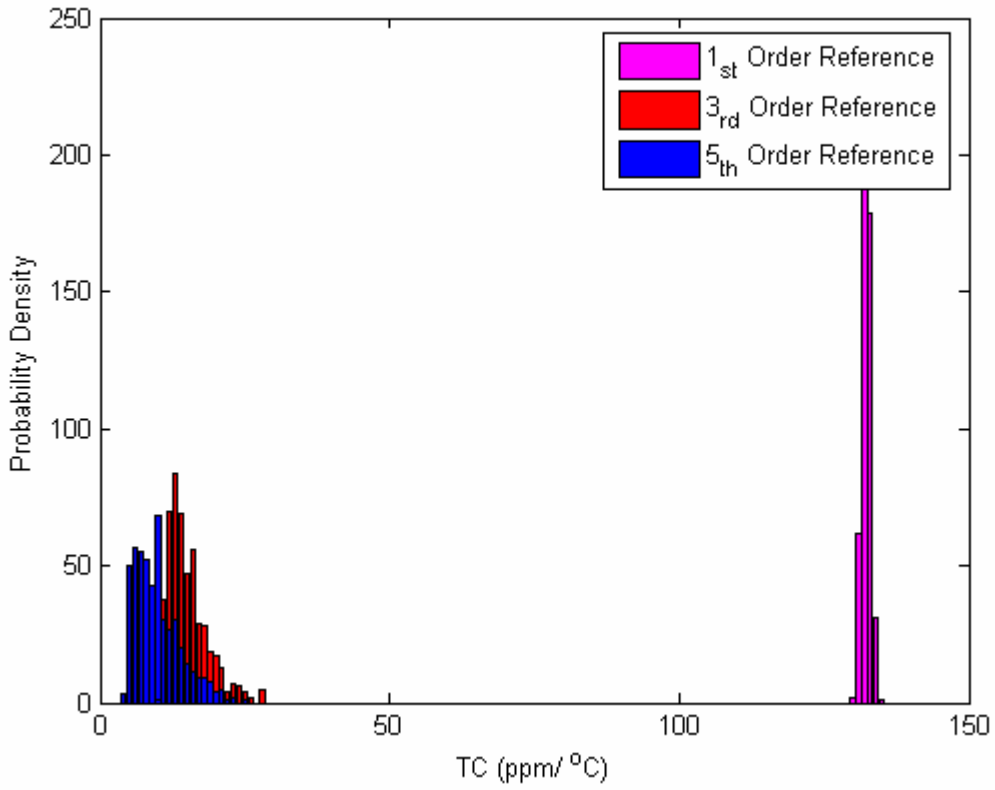

Figure 6.15: Process Variations of 1^{st}, 3^{rd}, and 5^{th} Order References, $-55^{o}C < T < 125^{o}C$

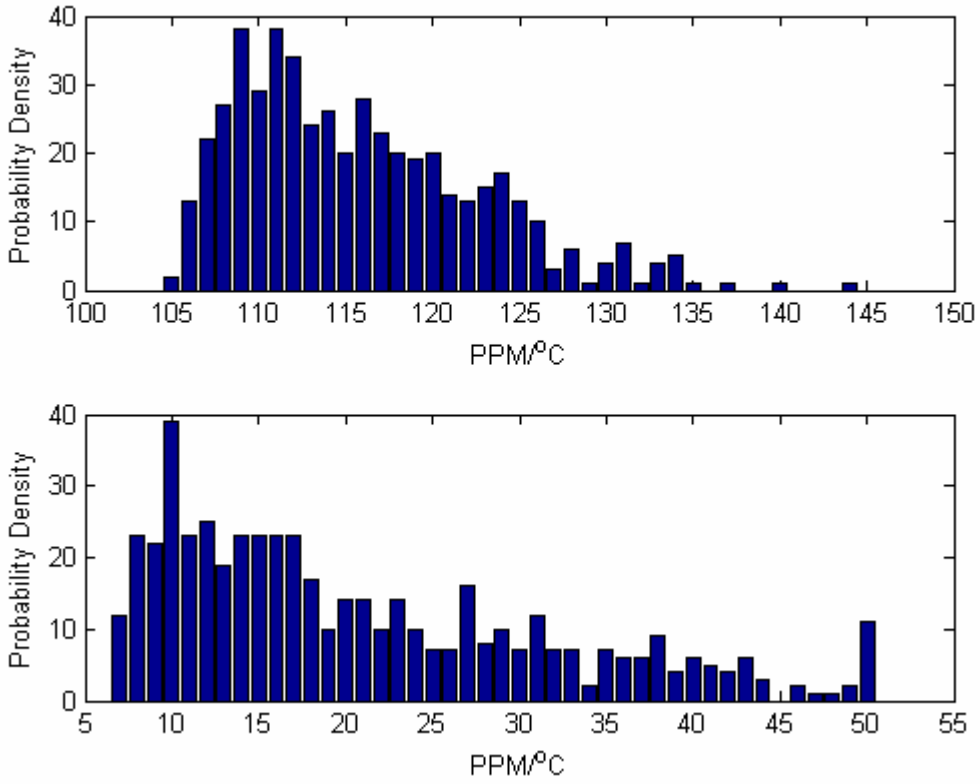

Figure 6.16: Process Variations of 1^{st} and 3^{rd} Order References, $0^o C < T < 70^o C$

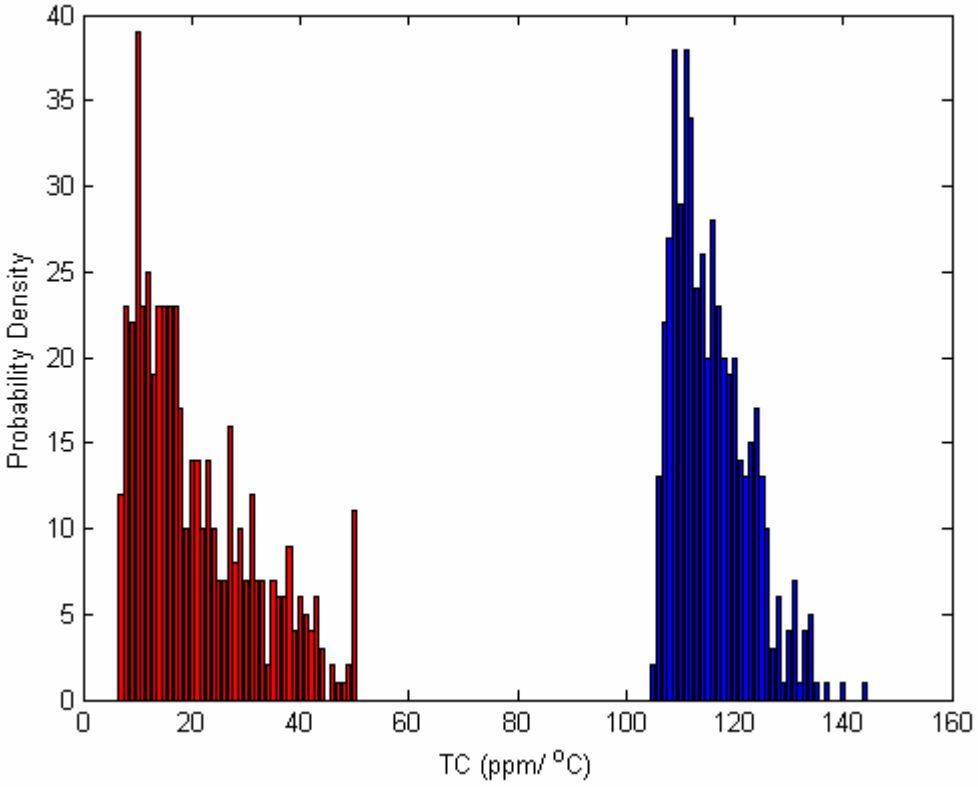

Figure 6.17: Process Variations of 1^{st} and 3^{rd} Order References, $0^o C < T < 70^o C$

Table 6.2: 1^{st} and 3^{rd} Order Process Variations of $90nm$, $0^o C < T < 70^o C$

1^{st} Order	$mean(V_{ref}) = 0.200112$ $\sigma(V_{ref}) = 1.51mV$ $mean(TC) = 116\ ppm/^o C$ $\sigma(TC) = 6.95\ ppm/^o C$
3^{rd} Order	$mean(V_{ref}) = 0.20005$ $\sigma(V_{ref}) = 1.38mV$ $mean(TC) = 20.92\ ppm/^o C$ $\sigma(TC) = 11.6\ ppm/^o C$

6.9 VOLTAGE REFERENCE DESIGN CONCLUSION

This chapter shows the performance capabilities of higher order voltage references in multiple technologies. Using the MLS algorithm, a solution for scaling the temperature dependent currents is found. With slight adjustments to the scaling coefficients given by the MLS algorithm to compensate for the non-ideal effects of the transistors, an optimized solution is completed.

The simulated performance results followed the expected results from the system analysis very closely. Transitioning from a first order to a third order voltage reference reduced the TC by approximately an order of magnitude. It was also demonstrated that going form a third order to a fifth order voltage reference approximately reduced the TC by half. Table 6.3 summarizes the non-process variation TC performances. The impact of process variation was reasonable and may be reviewed in Table 6.1 and Figure 6.11. However, from a 1^{st} order reference to a 5^{th} order reference, the same relative magnitudes of performance was maintained.

Table 6.3: TC of Voltage Reference Circuit without Process Variation

Technology	Order	Temp ^{o}C	TC $ppm/^{o}C$
0.18μ	1^{st}	$0^{o}C < T < 70^{o}C$	56
0.18μ	1^{st}	$-55^{o}C < T < 125^{o}C$	132
0.18μ	3^{rd}	$0^{o}C < T < 70^{o}C$	4.3
0.18μ	3^{rd}	$-55^{o}C < T < 125^{o}C$	11.3
0.18μ	5^{th}	$-55^{o}C < T < 125^{o}C$	4.8
$90nm$	1^{st}	$0^{o}C < T < 70^{o}C$	108
$90nm$	1^{st}	$-55^{o}C < T < 125^{o}C$	230
$90nm$	3^{rd}	$0^{o}C < T < 70^{o}C$	7.3

Chapter 7: Summary and Future Work

A voltage reference architecture that is able to scale with technology and perform with a low temperature coefficient is presented. The architecture is able to reduce the TC of the voltage reference by an order of magnitude compared to standard first order voltage references. This improvement allows the TC performance to compete with higher voltage bandgap references that are based on bipolar technologies. The circuit is comprised of CMOS devices and resistors, where there are no more than two CMOS devices stacked between the power supplies. This will guarantee the circuit operation to the lowest possible power supply voltage.

The reasons for scaling the dimensions of a CMOS device are compelling. With an ideal device, it is shown that an SOC's area and power consumption will decrease with the square of the scaling coefficient. Also, the SOC's digital frequency of operation and analog circuit bandwidth will gain directly proportionally to the scaling coefficient. These potentially rewarding goals are the motivation to scale the MOSFET devices.

Unfortunately, there is a price to pay for scaling the transistors. All of the electric field strengths are not preserved, as would occur with ideal models for the devices. Short channel effects of the devices force engineering trade-offs that hinder the performance of analog circuits. As discussed in Chapter 2, the V_{dd} to V_{th} ratio does not scale linearly. The ratio decreases, which reduces the amount of headroom available for analog circuits. For this reason, it was required to develop a circuit that uses only two devices stacked between V_{dd} and V_{ss} for future generations of technology.

As a result of the lost headroom and threshold voltage tradeoff, a balance between the power consumption of both digital circuits and analog circuits will have to

be maintained while trying to sustain the ideal scaling of I_{dsat} for frequency considerations. When a digital gate is powered off, it does not change states. In this condition, at least one device in the gate typically has V_{dd} as its V_{ds} voltage, so that there is not a short between the power supplies. The leakage current of this device is exponentially dependent on $V_{ds} - V_{th}$. Therefore, it is preferred for low power to have a higher threshold voltage at the expense of headroom. Other non-ideal effects are discussed in Chapter 2, which include dynamic range and power consumption. However, even with the non-ideal effects entering the scaling process, the manufactured devices are keeping up with the scaling trend and are reaping many of the benefits.

In order to gain insight into the operation of a temperature independent voltage reference, the fundamental temperature properties of silicon and their models are discussed in Chapter 3. Using the temperature dependent parameters of the CMOS device, CTAT and PTAT voltage and current references may easily be evaluated. Chapter 4 presented many types of PTAT and CTAT signal generating circuits that may be used to generate a temperature independent reference. All of the circuits have no more than two devices stacked between the power supplies, thus keeping the requirements of future generations of technology in mind and proving that there is more than one viable solution to this problem.

By using the basic temperature dependent signals from circuits like the ones shown in Chapter 4, a system that sums all of the temperature dependent signals together is shown to produce a temperature stable voltage reference. A methodology, introduced in Chapter 5, uses the idea of treating every temperature dependent parameter as a polynomial. By summing and scaling multiple temperature dependent polynomials together, a higher order voltage reference may easily be designed.

Applying the system theory discussed in Chapter 5 to actual circuits, several voltage references and their performance characteristics are shown in Chapter 6. A circuit design operating at 0.9 volts in a 0.18μ technology is developed. First order, third order, and fifth order circuit designs are presented. In all cases, the circuit's temperature response improves as predicted by the theory presented in Chapter 5. In order to show the scalability of this architecture, first and third order solutions are also presented in a $90nm$ process operating at 0.4 volts. The results show that a fully CMOS scalable voltage reference is a viable option for the SOCs and their future technologies.

The precedent has been set that a solution for a scalable, high performing, fully CMOS voltage reference. The circuits proposed are targeted to be manufactured within the next year, pending the availability and scheduling of the shuttles for the 0.18μ and $90nm$ process. Additional work that would be beneficial, incorporates a circuit to extract the threshold voltage. With this additional information, a model of the threshold voltage's response to temperature may be verified and updated, if required.

Additional work includes improvements for manufacturing by reducing the architecture's sensitivity to mismatching of devices and resistors. One alternative is to convert the architecture into a switched capacitor circuit. A good starting point would be to switch two different currents into a single diode-connected device to generate the ΔV_{gs} that is responsible for the PTAT voltage. This concept can be expanded throughout the voltage reference circuit.

Bibliography

[1] Anne-Johan Annema, Bram Nauta, Ronald van Langevelde, and Hans Tuinhout, "Analog Circuits in Ultra-Deep-Submicron CMOS," in IEEE Journal of Solid-State Circuits, Vol. 40, No. 1, January 2005, pp. 132-143.

[2] Bill Grundmann, Rajesh Galivanche, and Sandip Kundu, "Circuit and Platform Design Challenges in Technologies beyond 90nm," in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, 2003.

[3] Kerry Bernstein, Ching-Te Chuang, Rajiv Joshi, and Ruchir Puri, "Design and CAD Challenges in sub-90nm CMOS Technologies," in ICCAD, November 11-13, 2003, pp. 129-136.

[4] Arijit Raychowdhury, Sailbal Mukhopadhyay, and Kaushik Roy, "Modeling and Estimation of Leakage in Sub-90nm Devices," in Proceedings of the 17th International Conference on VLSI Design (VLSID'04), 2004.

[5] Ban P. Wong, Anurag Mittal, Yu Cao, and Greg Starr, Nano-CMOS Circuit and Physical Design. New Jersey: John Wiley & Sons, Inc., 2005.

[6] Pervez M. Aziz, Henrik V. Sorensen, and Jan Van Der Spiegel, "An Overview of Sigma-Delta Converters," in IEEE Signal Processing Magazine, January 1996, pp. 61-84.

[7] Eric Soenen, "Analog Circuit Design in Deep Submicron CMOS Processes," in Deep-Submicron Analog and RF Circuit Design ISSCC 2004 Short Course.

[8] Trond Ytterdal, Yuhua Cheng, and Tor A. Fjeldly, Device Modeling for Analog and RF CMOS Circuit Design, West Sussex, England: John Wiley & Sons Ltd., 2003.

[9] Scott Thompson, Paul Packan, and Mark Bohr, "MOS Scaling: Transistor Challenges for the 21st Centery," Intel Tecnology Journal, Q3 1998.

[10] Yuan Taur, et. al., "CMOS Scaling into the Nanometer Regime," Proceedings of the IEEE, Vol. 85, No. 4, April 1997.

[11] S. M. Sze, Physics of Semiconductor Devices, second edition, New York: John Wiley & Sons, Inc., 1981.

[12] Yannis Tsividis, Operation and Modeling of The MOS Transistor, second edition, Boston: WCB/McGraw-Hill, 1999.

[13] E. J. Nowak, "Maintaining the Benefits of CMOS Scaling When Scaling Bogs Down," in IBM J. Res. & Dev., Vol. 46, No. 2/3 March/May 2002.

[14] Kai Chen, et. al., "Accurate Models for CMOS Scaling and Gate Delay in Deep Sub-micron Regime," in IEEE, 1997.

[15] Behzad Razavi, Design of Analog CMOS Integrated Circuits, New York: McGraw-Hill, 2001.

[16] Robert h. Dennard, et. al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE Journal of Solid-State Circuits, Vol. sc-9, No. 5, October 1974.

[17] Robert F. Pierret, Field Effect Devices Volume IV, second edition, Massachusetts: Addison-Wesley Publishing Company, 1990.

[18] Gerold W. Neudeck, The PN Junction Diode Volume II, second edition, Massachusetts: Addison-Wesley Publishing Company, 1990.

[19] R. Gwoziecki, T. Skotnicki, "Physics of the Subthreshold slope – initial improvement nad final degradation in short CMOS devices," in ESSDERC, 2002.

[20] Kuan-Yu Fu and Yuk L. Tsang, "On the Punchthrough Phenomenon in Submicron MOS Transistors," IEEE Transactions on Electron Devices, Vol. 44, No. 5, May 1997.

[21] Kaushik Roy, et. al., "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," Proceedings of the IEEE, Vol. 91, No. 2, February 2003.

[22] K. M. Cao, et. al., "BSIM4 Gate Leakage Model Including Source-Drain Partition," IEEE 2000.

[23] W. C. Lee and C. Hu, " Modeling gate and substrate currents due to conduction and valence band electron and hole tunneling," Proceedings of the Symposium on VLSI Technology, pp. 198-199, 2000.

[24] Aldert van der Ziel, Noise in Solid State Devices and Circuits, Canada: John Wiley & Sons, Inc., 1986.

[25] Jung-Suk Goo, et. al., "Physical Origin of the Excess Thermal Noise in Short Channel MOSFETs," IEEE Electron Device Letters, Vol. 22, No. 2, February 2001.

[26] Dimitris P. Triantis, et. al., "Thermal Noise Modeling for Short-Channel MOSFET's," IEEE Transactions on Electron Devices, Vol. 43, No. 11, November 1996.

[27] Gerhard Knoblinger, et. al., "A new Model for Thermal Channel Noise of Deep Submicron MOSFET's and its Application in RF-CMOS Design," VLSI 2000.

[28] Gerhard Knoblinger, et. al., "Thermal Channel Noise of Quarter and Sub-Quarter Micron NMOS FET's," ICMTS 2000.

[29] A. A. Abidi, " High-Frequency Noise Measurements on FET's with Small Dimensions," IEEE Transactions on Electron Devices, Vol. ED-33, No. 11, November 1986.

[30] O. M. Bulashenko and J. M. Rubi, "Excess Noise Caused By Transverse Inhomogeneity of Conductive Channels," Applied Physics Letters, Volume 73, No. 2, July 13, 1998.

[31] F. Bonani and G. Ghione, "Noise in Semiconductor Devices, Modeling and Simulation," Germany: Springer-Verlag Berlin Heidelberg, 2001.

[32] "Process Integration, Devices, and Structures," The International Technology Roadmap for Semiconductors (ITRS), 2004.

[33] Thomas H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, New York: Cambridge University Press, 1998.

[34] Fariborz Assaderaghi, et. al., "Dynamic Threshold-Voltage MOSFET (DTMOS) for Ultra-Low Voltage VLSI," IEEE Transactions On Electron Devices, Vol. 44, No. 3, March 1997.

[35] Bijan Davari, et. al. "CMOS Scaling for High Performance and Low Power – The Next Ten Years," Proceedings of the IEEE, Vol 83., No. 4, April 1995.

[36] Igor Polishchuk and Chenming Hu, "Polycrystalline silicon/metal stacked gate for threshold voltage control in metal-oxide-semiconductor field-effect transistors," Applied Physics Letters, Volume 76, Number 14, April 3, 2000.

[37] Valarie Welsh, " Work Function Engineering With Molybdenum and Molybdenum-Nitride Gate PMOS," 22nd Annual Microelectronic Engineering Conference, May 2004.

[38] Xiangli Li, et. al., "Threshold Voltage Control for Deep Sub-micrometer Fully Depleted SOI MOSFET," Proceedings of the 15th Biennial University/Government/Industry Microelectronics Symposium 2003.

[39] Rafael Rios and Narain D. Arora, "Determination of Ultra-Thin Gate Oxide Thicknesses for CMOS Structures Using Quantum Effects," IEEE International Electron Devices Meeting (IEDM), 1994.

[40] E. Conwell and V. F. Weisskopf, "Theory of Impurity Scattering in Semiconductors," Physical Reiview, Volume 77, Number 3, February 1, 1950.

[41] Ben G. Streetman, Solid State Electronic Devices, 3rd edition, New Jersey: Prentice-Hall, Inc., 1990.

[42] Frank Stern and W. E. Howard, "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit," Physical Review, Volume 163, Number 3, November 15, 1967.

[43] Y. P. Varshni, "Temperature Dependence Of The Energy Gap In Semiconductors," Physica, Volume 34, pg. 147-154, 1967.

[44] Frank Herman, "The Electronic Energy Band Structure of Silicon and Germanium," Proceedings of the IRE, December, 1955.

[45] C. D. Thurmond, "The Standard Thermodynamic Functions for the Formation of Electrons and Holes in Ge, Si, GaAs, and GaP," J. Electrochem. Soc.: Solid-State Science and Technology, August, 1975.

[46] R. Passler, "Parameter Sets Due to Fittings of the Temperature Dependencies of Fundamental Bandgaps in Semiconductors," Phys. Stat. Sol. (b), 1999.

[47] Frank Herman, "Theoretical Investigation of the Electronic Energy Band Structure of Solids," Reviews of Medern Physics, Vol. 30, No. 1, January, 1958.

[48] Benjamin Lax, "Experimental Investigations of the Electronic Band Structure of Solids," Reviews of Modern Physics, Vol. 30, No. 1, January, 1958.

[49] Robert F. Pierret and Gerold W. Neudeck, Semiconductor Fundamentals Volume I, second edition, Massachusetts: Addison-Wesley Publishing Company, 1990.

[50] Smartspice BSIM3 or proper manual.

[51] Theodore L. Brown and H. Eugene LeMay, Jr., Chemistry: The Central Science, 4th edition, New Jersey: Prentice Hall, 1988.

[52] J. Bardeen and W. Shockley, "Scattering of Electrons in Crystals in the Presence of Large Electric Fields," Physical Review, Volume 80, No. 1, October 1, 1950.

[53] P. P. Debye and E. M. Conwell, "Electrical Properties of N-Type Germanium," Physical Review, Volume 93, No. 4, February 15, 1954.

[54] S. M. Sze and J. C. Irvin, "Resistivity, Mobility, and Impurity Levels In GaAs, Ge, and Si at 300K," Solid-State Electonics, Vol. 11, 1968, pp. 599-602.

[55] C. Jacoboni, et. al., "A Review of Some Charge Transport Properties of Silicon," Solid-State Electronics, Vol. 20, 1977, pp. 77-89.

[56] Eric Vittoz and Jean Fellrath, "CMOS Analog Integrated Circuits Based on Weak Inversion Operation," IEEE Journal of Solid-State Circuits, Vol. SC-12, No. 3, June 1997.

[57] Ana Isabela Araujo Cunha, et. al., "An Explicit Physical Model for the Long-Channel MOS Transistor Including Small-Signal Parameters," Solid-State Electronics, Vol. 38, No. 11, 1995.

[58] Matthias Bucher, et. al., "The EPFL-EKV Model Equations for Simulation," Technical Report, Model Version 2.6, June 1997, http://legwww.epfl.ch/ekv/.

[59] Christian C. Enz, etl al., "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications," Analog Integrated Circuits and Signal Processing, Vol. 8, pp. 83-114, 1995.

[60] Ana Isabela Araujo Cunha, et. al., "An MOS Transistor Model for Analog Circuit Design," IEEE Journal of Solid-State Circuits, Vol. 33, No. 10, October 1998.

[61] MOS Modeling short course from ISSCC Feb. 2004.

[62] Ghong-Gun Yu and Randall L. Geiger, "An Accurate and Matching-Free Threshold Voltage Extraction Scheme for MOS Transistors," IEEE International Symposium on Circuits and Systems (ISCAS), 1994.

[63] David F. Mietus, "Reference Voltage Circuit Having A Substantially Zero Temperature Coefficient," United States Patent #5,666,046, 1997.

[64] Gabriel A. Rincon-Mora and Phillip E. Allen, "A 1.1-V Current-Mode and Piecewise-Linear Curvature-Corrected Bandgap Reference," IEEE Journal of Solid-State Circuits, Vol. 33, No. 10, October 1998.

[65] Randal Geiger, Iowa State University, personal correspondence.

[66] E. M. Camacho-Galeano, et. al. ,"A 2-nW 1.1-V Self-Biased Current Reference in CMOS Technology," IEEE Transactions on Circuits and Systems II, Volume 52, No. 2, February 2005.

[67] Erwin Kreyszig, Advanced Engineering Mathematics, 6th edition, New York: John Wiley & Sons, Inc., 1988.

[68] Perry Miller and Doug Moore, "Precision Voltage References," Texas Insruments Analog Applications Journal, November 1999.

[69] Gabriel Alfonso Rincon-Mora, Voltage References: From Diodes to Precision High-Order Bandgap Circuits, New York: John Wiley & Sons, Inc., 2002.

[70] Peter R. Kinget, "Device Mismatch and Tradeoffs in the Design of Analog Circuits," IEEE Journal of Solid-State Circuits, Vol. 40, No. 6, June 2005.

[71] D. F. hilbiber, "A New Semiconductor Voltage Standard," IEEE International Solid-State Circuits Conference, February 19, 1964.

[72] R. J. Widlar, "New Developments in IC Voltage Regulators," IEEE International Solid-State Circuits Conference, February 20, 1970.

[73] R. J. Widlar, "New Developments in IC Voltage Regulators," IEEE Journal of Solid-State Circuits, Vol. sc-6, No. 1, February 1971.

[74] Paul R. Gray and Robert G. Meyer, Analysis and Design of Analog Integrated Circuits, second edition, New York: John Wiley & Sons, Inc. 1984.

[75] Yannis P. Tsividis, "Accurate Analysis of Temperature Effects in $I_c - V_{be}$ Characteristics with Application to Bandgap Reference Sources," IEEE Journal of Solid-State Circuits, Vol. SC-15, No. 6, December 1980.

[76] Hironori Banba, et. al., "A CMOS Bandgap Reference Circuit with Sub-1-V Operation," IEEE Journal of Solid-State Circuits, Vol. 34, No. 5, May 1999.

[77] Andrea Pierazzi, Andrea Boni, and Carlo Morandi, "Band-Gap References for near 1-V operation in standard CMOS technology," IEEE 2001 Custom Integrated Circuits Conference.

[78] Ka Nang Leung and Philip K. T. Mok, "A Sub-1-V 15-ppm/^oC CMOS Bandgap Voltage Reference Without Requiring Low Threshold Voltage Device," IEEE Journal of Solid-State Circuits, Vol. 37, No. 4, April, 2002.

[79] K. Lasanen, et. al., "Design of a 1-V Low Power CMOS Bandgap Reference Based On Resistive Subdivision," Proceedings of the 45th IEEE Midwest Symposium on Circuits and Systems, Tulsa, Oklahoma, August 2002.

[80] Arne E. Buck, et. al., "A CMOS Bandgap Reference Without Resistors," IEEE Journal of Solid-State Circuits, Vol. 37, No. 1, 2002.

[81] Francisco Serra-Graells and Jose Luis Huertas, "Sub-1-V CMOS Proportional-to-Absolute Temperature References," IEEE Journal of Solid-State Circuits, Vol. 38, No. 1, January 2003.

[82] Ka Nang Leung and Philip K. T. Mok, "A CMOS Voltage Reference Based on Weighted ΔV_{GS} for CMOS Low-Dropout Linear Regulators," IEEE Journal of Solid-State Circuits, Vol. 38, No. 1, January 2003.

[83] S. Sengupta, et. al., "A Process, Voltage, and Temperature Compensated CMOS Constant Current Reference," IEEE ISCAS, 2004.

[84] Song Liu and R. Jacob Baker, "Process and Temperature Performance of a CMOS Beta-Multiplier Voltage Reference," IEEE, 1999.

[85] Andrea Boni, "Op-Amps and Startup Circuits for CMOS Bandgap References With Near 1-V Supply," IEEE Journal of Solid-State Circuits, Vol 37, No. 10, October 2002.

[86] Gerhard Knoblinger, et. al., "A new Model for Thermal Channel Noise of Deep Submicron MOSFET's and its Application in RF-CMOS Design," VLSI 2000.

[87] Dimitris P. Triantis, et. al., "Thermal Noise Modeling for Short-Channel MOSFET's," IEEE Transactions on Electron Devices, Vol. 43, No. 11, November 1996.

[88] Jung-Suk Goo, et. al., "Physical Origin of the Excess Thermal Noise in Short Channel MOSFETs," IEEE Electron Device Letters, Vol. 22, No. 2, February 2001.

[89] Marcel J. M. Pelgrom, et. al., "Matching Properties of MOS Transistors," IEEE Journal of Solid-State Circuits, Vol. 24, No. 5, October 1989.

[90] Jerome Dubois, et. al., "Impact of Source/Drain Implants on Threshold Voltage Matching in Deep Sub-Micron CMOS Technologies," ESSDERC 2002.

[91] Kadaba R. Lakshmikumar, et. al., "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design," IEEE Journal of Solid-State Circuits, Vol. sc-21, No. 6, December 1986.

[92] P. R. Kinget, "Device Mismatch and Tradeoffs in the Design of Analog Circuits," IEEE Journal of Solid-State Circuits, Vol. 40, No. 6, June 2005.

Vita

Michael David Cave was born in Mesa, Arizona in 1969. He completed his B.S.E.E. degree at the University of Arizona in 1992 and moved to Austin, Texas to join Motorola, Inc. While working as an analog circuit design engineer, he completed his M.S.E. from the University of Texas at Austin. From 1997 until 2001, he continued his career as an analog circuit and systems design engineer as a senior member of the technical staff at Sigmatel, Inc. From 2001 to 2004 he worked at ViXS Systems, Inc., as a senior member of the technical staff, until he resigned to pursue his Ph.D. full time. The technical experience he has gained, both from academics and industry, is fairly broad. He has been involved in the circuit design of ADCs, DACs, RF transceivers, IR transceivers, PLLs, VCXOs, references, and OFDM systems (including ADSL and 802.11a). Most of his experience resides in integrated circuit design with some board level design. However, he gained communication theory experience when developing the communication chip sets associated with his previous employments. Michael Cave has been elected as a Senior Member of IEEE and has authored several publications in the areas of analog circuits, communication systems, and low power analog circuit design. He is an inventor on 10 patents, with additional patents pending.

Permanent address: 3106 Barton Point Circle, Austin, Texas, 78733

This dissertation was typed by the author.